PREDIT

SILVERPOLIS

Modélisation de la tarification sensible à l'usage dans les modèles mésoscopiques: étude comparative

Rapport final

Octobre, 2012

Ecole Normale Supérieure de Cachan (ENS Cachan)

Royal Institute of Technology (KTH)

Un projet du PREDIT 4, GO 6 : Politiques de transports

Dans le cadre du programme européen ERA-NET, ENT 17 SURPRICE.

Ecole Normale Supérieure de Cachan (ENS Cachan)
Département d'Economie et Gestion
61 avenue du président Wilson
94320 Cachan, FRANCE.
Contact : Professeur André de Palma, Courriel: andre.depalma@ens-cachan.fr

Royal Institute of Technology (KTH)
Centre for Transport Studies
Teknikringen 14
10044 Stockholm, SWEDEN.
Contact : Dr. Leonid Engelson, Courriel: lee@abe.kth.se

Equipe:

ENS Cachan

André de Palma, ENS Cachan, France (Coordinateur)

Mohammad Saifuzzaman, ENS Cachan, France,

KTH

Leonid Engelson, KTH, Stockholm, Sweden (Coordinateur)

Ida Kristoffersson, KTH, Stockholm, Sweden

Table of Contents

Remerciements

Nous tenons à remercier celles et ceux qui nous ont accompagnés dans l'élaboration de ce projet. En premier lieu les Professeurs Mogens Fosgerau de l'Université technique de Danemark et Robin Lindsey de l'Université de Colombie britannique (Canada) en tant que coauteurs des contributions théoriques. Ensuite, nous remercions la DRAST, en particulier Monsieur Gérard Brun. Egalement les participants du Carrefour de PREDIT pour leurs remarques et leurs propositions. L'organisation de la réunion des utilisateurs de METROPOLIS est effectuée avec le concours de l'Université technique de Berlin dans le cadre du projet européen SustainCity.

Modélisation de la tarification sensible à l'usage dans les modèles mésoscopiques: étude comparative

Introduction

Après Londres, Stockholm était la deuxième agglomération européenne à instaurer un péage urbain. Cette expérience a fait l'objet d'études détaillées sous des points de vue variés afin de mieux comprendre les différents aspects socio-économiques de cette politique de transports. Le projet SilverPolis vise à faire avancer l'état de l'art dans le domaine de la modélisation dynamique de trafic en se basant sur les données obtenues par cette expérience. Dans le cadre de ce projet nous passons d'abord en revue les modèles utilisés pour étudier le péage urbain, puis de manière plus détaillée, nous nous concentrons sur deux modèles METROPOLIS et SILVESTER. La comparaison se focalise sur les aspects de conception du système, de traitement de la congestion et de modélisation des effets environnementaux et économiques.

Ces modèles sont tous deux dynamiques et prennent en compte le choix de l'heure de départ et de la route. Cependant, leur base théorique et leur structure diffèrent. Par ailleurs, SILVESTER est déjà calé et validé pour le réseau de Stockholm avec les données issues de l'expérience de péage urbain de 2006. METROPOLIS doit être appliqué et calé pour ce réseau. Ensuite les données d'expérience de péage doivent y être introduites et les résultats obtenus seront comparés.

L'objectif du projet consiste à comprendre la relation entre les propriétés des modèles et la qualité des prévisions. Il analyse les conditions requises pour le développement des modèles afin de prédire les effets de mise en place d'une politique de péage urbain. Dans le cadre de ce projet, nous souhaitons identifier les caractéristiques clés pour obtenir des prévisions raisonnables en matière de conséquences et de retombées du péage urbain. Cela pourrait faciliter le développement d'une plateforme européenne d'outil d'aide à la décision en matière de tarification des réseaux urbains européens.

En particulier, nous nous proposons de répondre aux questions suivantes:

1. Quelles sont les différences les plus marquantes entre les prédictions de METROPOLIS et de SILVESTER en termes de conséquences de la mise en place du péage à Stockholm ?
2. Quelles sont les différences les plus importantes entre les effets prédits du péage à Stockholm et à Paris ?
3. Quelles sont les caractéristiques clés d'un modèle pour assurer les prédictions raisonnables des effets d'une politique de péage urbain ?

Articles de recherche publies ou soumis à publication

A. Dynamic and Static Congestion Models

Reference:

de Palma, A. and Fosgerau, M., (2011), Dynamic and Static Congestion Models: a Review. In *Handbook in Transport Economics*, A. de Palma, R. Lindsey, E. Quinet et R. Vickerman, (eds.), Edgar Elgard.

Résumé

Nous commençons par une revue de l'approche conventionnelle de l'équilibre statique. Dans un tel modèle les flux des déplacements et les délais dus à la congestion sont supposés d'être constants dans le temps. Le modèle statique ne spécifie pas l'intervalle du temps durant lequel chaque déplacement se réalise. Cela constitue l'un des inconvénients parce que le modèle statique n'est pas capable de décrire l'évolution de la durée de congestion en fonction de modification de la demande ou de la capacité des routes. Le modèle de goulot d'étranglement de Vickrey / Arnott, de Palma et Lindsey (1992) résout ce problème. Ce modèle associe la congestion, sous forme des queues développées derrière des goulots d'étranglement, avec les préférences et le choix des usagers en matière de l'heure de départ. Nous présentons l'équilibre des usagers et l'optimum social dans ce modèle et nous expliquons comment l'optimum peut être décentralisé par application d'un péage variable dans le temps. Ensuite, quelques extensions de ce modèle sont présentées telles que prise en compte de la demande élastique, de l'hétérogénéité des usagers, de la demande et de la capacité stochastique et l'application aux petits réseaux. Nous concluons par l'identification de quelques problèmes non-résolus de modélisation qui s'appliquent non seulement au modèle de goulot d'étranglement mais également qu'aux préférences des usagers dans le choix de l'heure de départ et à la dynamique de la congestion en général.

Abstract

We begin by providing an overview of the conventional static equilibrium approach. In such model both the flow of trips and congestion delay are assumed to be constant. A drawback of the static model is that the time interval during which travel occurs is not specified so that the model cannot describe changes in the duration of congestion that result from changes in demand or capacity. This limitation is overcome in the Vickrey/Arnott, de Palma Lindsey bottleneck model, which combines congestion in the form of queuing behind a bottleneck with users' trip-timing preferences and departure time decisions. We derive the user equilibrium and social optimum for the basic bottleneck model, and explain how the optimum can be decentralized using a time-varying toll. They then review some extensions of the basic model that encompass elastic demand, user heterogeneity, stochastic demand and capacity and small networks. We conclude by identifying some unresolved modelling issues that apply not only to the bottleneck model but to trip-timing preferences and congestion dynamics in general

B. Traffic Congestion Pricing Methodologies and Technologies

Reference:

de Palma, A. and Lindsey, R., (2011), Traffic Congestion Pricing Methodologies and Technologies, *Transportation Research Part C: Emerging Technologies*, 19(6), 1377-1399.

Résumé

Ce travail synthétise les méthodes et les technologies de la tarification routière. Le péage peut être instauré à différents niveaux allant d'une seule voie de circulation jusqu'à un tronçon routier ou l'ensemble du réseau national. Les tarifs peuvent varier en fonction de l'heure de la journée, type de route ou les caractéristiques de la voiture ou même en temps réel par rapport à la situation de circulation. Les stations traditionnelles de péage ont aujourd'hui laissé la place aux systèmes électroniques modernes de péage. Les principales technologies utilisées sont divisées en deux catégories : (a) Les systèmes installés au bord de la route qui reconnaissent les véhicules par photographie de la plaque d'immatriculation ou par l'utilisation des badges qui communiquent avec une antenne dédiée par la micro-onde. (b) Les systèmes embarqués dans la voiture qui communiquent à l'aide de satellite ou de réseau de téléphonie mobile. La rapidité d'instauration du péage et son étendue dépendent de la technologie choisie ainsi que d'autres fonctionnalités et services qui en sont attendus.

Abstract

This paper reviews the methods and technologies for congestion pricing of roads. Congestion tolls can be implemented at scales ranging from individual lanes on single links to national road networks. Tolls can be differentiated by time of day, road type and vehicle characteristics, and even set in real time according to current traffic conditions. Conventional toll booths have largely given way to electronic toll collection technologies. The main technology categories are roadside-only systems employing digital photography, tag and beacon systems that use short-range microwave technology, and in vehicle-only systems based on either satellite or cellular network communications. The best technology choice depends on the application. The rate at which congestion pricing is implemented, and its ultimate scope, will depend on what technology is used and on what other functions and services it can perform.

C. Network market conduct with atomic and non-atomic players

Refrence:

Engelson, L., I. Kristoffersson, A. de Palma, K. (2012), Network market conduct with atomic and non-atomic play. Working paper.

Résumé

On considère un jeu de Stackelberg dans un réseau statique constitué par deux routes en parallèle et deux types d'usagers : il y un continuum de conducteurs et une flotte de camions. Les fonctions de congestion sont affines et spécifiques au groupe. Chaque voiture est suffisamment petite (atome) pour ignorer l'impact de son choix de route sur le niveau de congestion. Le coordinateur de la flotte de camions peut prédire l'impact de ses choix de route sur le niveau de congestion de sa flotte. On considère quatre scénarios : le coordinateur réduit le coût de trajet des véhicules de sa flotte, l'optimum social, l'optimum de second rang lorsque le coordinateur tente de minimiser le coût social pour l'ensemble des véhicules (voitures et camions), ainsi que le cas de référence, qui correspond à une absence de coordination. Nous montrons que les solutions du premier scénario coïncident avec l'équilibre de Wardop lorsque le nombre de camions est suffisamment petit ; le coût peut augmenter ou diminuer, lorsque le nombre de camions est élevé. L'ensemble des solutions intérieures n'est pas plus petit que dans le cas de l'équilibre de Wardop, mais n'est pas plus grand que pour l'équilibre de seconde espèce. Finalement nous montrons que le nombre de camions sur une route peut être une fonction non monotone du nombre total de camions. On peut aussi observer plusieurs discontinuités dans le nombre d'usagers quand le nombre total de camions augmente.

Abstract

We consider a Stackelberg game in a static network with two routes in parallel and two user groups: a continuum of cars and a fleet of trucks. The congestion functions are affine and group specific. Each car is non-atomic and ignores the impact of his route choice on congestion. On the contrary, the coordinator of the trucks can predict the response of cars to her routing strategies. We consider several scenarios: the coordinator reducing the total travel cost of the trucks, the social optimum, the second-best optimum with coordinator attempting to minimize the total system cost, as well as the benchmark, with no coordination at all. We show that solution to the first scenario coincides with user equilibrium for small number of trucks and may improve or worsen it for large number of trucks. The set of interior solutions for the first scenario is not less than in the user equilibrium and not larger than in the second best. Finally, the route usage by trucks can be non-monotonic and perform multiple jumps when the size of the fleet increases.

D. Comparison of two dynamic transportation models: The case of Stockholm congestion charging

Refrence:

Engelson, L., I. Kristoffersson, A. de Palma, K. Motamedi, and M. Saifuzzaman, (2012), Comparison of two dynamic transportation models: The case of Stockholm congestion charging. In *proceedings of the 4th TRB Conference on Innovations in Travel Modeling*, Florida, USA.

Résumé

Cet article passe en revue des modèles de transport utilisés pour la prédiction des effets de la tarification contre la congestion routière dans le cas des villes européennes et effectue une comparaison approfondie de deux de ces modèles, à savoir METROPOLIS et SILVESTER. Ils sont tous les deux des modèles dynamiques et mésoscopics qui prennent en compte le choix modal, le choix de route et celui de l'heure de départ. Ils sont callés pour la ville de Stockholm avant et après la mise en place de tarification. Les résultats obtenus par des deux modèles sont comparés entre eux et validés par rapport aux observations des effets de la tarification en Stockholm. Tous les deux modèles apportent une amélioration significative dans le réalisme des résultats par rapport aux modèles statiques. Néanmoins, les résultats d'analyse coût-bénéfice selon deux modèles peuvent être sensiblement différents.

Abstract

This paper reviews the transportation models used for predicting impacts of congestion charging in European cities and carries out in-depth comparison of two such models, METROPOLIS and SILVESTER. Both are mesoscopic dynamic models involving modal split, route choice and departure time choice calibrated for the Stockholm baseline situation without charges and applied for modeling effects of congestion charging. The results obtained from the two models are mutually compared and validated against actual outcome of the Stockholm congestion charging scheme. Both models provide significant improvement in realism over static models. However results of cost benefit analysis may differ substantially.

Modelling Impacts of Road User Charging with Mesoscopic Models: Comparative Study

1. Introduction

Traffic congestion is common in large cities and on major highways and it imposes a significant burden in lost time, uncertainty, and aggravation for passenger and freight transportation. Most of the costs of traffic congestion are borne by travellers collectively, but because individual travellers impose delays on others they do not pay the full marginal social cost of their trips and therefore create a negative externality. The standard economic prescription to internalize the costs of a negative externality is a Pigouvian tax. In the first edition of his textbook, *The Economics of Welfare*, Pigou (1920) himself argued for a tax on congestion and thereby launched the literature on congestion pricing. Congestion pricing has a big advantage over other transportation demand management policies in that it encourages travellers to adjust all aspects of their behaviour: number of trips, destination, mode of transport, time of day, route, and so on, as well as their long-run decisions on where to live, work and set up business.

For decades congestion pricing remained largely an ivory-tower idea, but interest gradually spread outside academia and congestion pricing has come into limited practice. The main operating schemes are High Occupancy Toll (HOT) lane facilities in the US, the London congestion charge, the Stockholm cordon charge, and Singapore's Electronic Road Pricing system. Several countries have considered regional or national road-pricing schemes — in part to internalize congestion and other traffic externalities. However, despite the apparent success of existing schemes, and plans to establish more, congestion pricing continues to be a hard sell. Several major proposals have recently been scuttled by public or political opposition. These setbacks illustrate the difficulties of designing congestion pricing schemes that are both efficient and publicly acceptable.

There is a large scientific literature available on impacts of congestion charging (Pigou, 1920; Vickrey, 1969; Small, 1983; Arnott et al., 1994; Glazer and Niskanen, 2000). The literature considering modeling of congestion charging is however more limited, e.g. Koh and Shepherd (2006). In practice, static assignment models integrated with travel demand models are often applied to forecast the impact in feasibility studies of congestion charging. This has been the case for example in Oslo (Odeck et al., 2003), Stockholm (Eliasson and Mattsson, 2006) and Copenhagen (Rich and Nielsen, 2007; Nielsen et al., 2002). It has however been agreed in the research community that the temporal aspects of congestion have a crucial role on system level. For example, the forecasts made with static models for Stockholm congestion charging system resulted in severe overestimation of impact on traffic flows during the peak hour and, at the same time, great underestimation of changes in travel times (Engelson and van Amelsfort, 2011). Moreover, the most effective charges aim to redistribute trips in time in order to cut down the congestion peak. Therefore impact of time-varying charges on departure time choice is an important issue. A mesoscopic dynamic model (MDM) can capture the time-varying aspect of congestion and congestion charging. At the same time it is not as detailed as a microscopic model. A mesoscopic assignment model integrated with a travel demand model is therefore suitable for calibration of whole city networks and thus for modeling impacts of city-wide congestion charging schemes. For a recent survey of dynamic models we refer the reader to de Palma and Fosgerau (2011).

METROPOLIS (de Palma et al., 1997) and SILVESTER (Kristoffersson and Engelson, 2009) are two state-of-the-art MDMs developed in the last decade with specific focus on congestion charging applications. De Palma et al. (2005) analyze different congestion charging schemes using METROPOLIS and a stylized urban road network. Marchal and de Palma (2002) apply METROPOLIS to Paris, and also give guidelines for model designers and planners who consider a shift to dynamic traffic simulation. Using METROPOLIS de Palma and Lindsey (2006) assess phase implementation of charging in Paris. SILVESTER is applied to Stockholm in Kristoffersson (2011). Kristoffersson and Engelson (2011) use SILVESTER to evaluate efficiency and equity of alternative congestion charging schemes for Stockholm.

There are very few opportunities to validate transportation models by observed response to charging. In Stockholm we have the unique possibility to use measurements from the field to validate transport models. Therefore both SILVESTER and METROPOLIS are in this study calibrated to Stockholm conditions in the situation without charging. Model response to the charges are then compared both between the two transport models and to measurements; this in order to provide a benchmark for modelling of congestion charging and in order to find model properties that are important for correct prediction. A similar in-depth comparative study of transportation models suitable for predicting impacts of congestion charging has to our knowledge not been undertaken before. Given that METROPOLIS and SILVESTER share the same ambition to improve conventional static transportation modelling of impacts of congestion charging by using dynamic modelling, but approaches the task in different ways, there is a good opportunity to compare implications of different modelling strategies.

2. State of practice in modelling of congestion charges

2.1. Methods of congestion pricing

Congestion pricing schemes can be categorized along several dimensions: (1) the type of scheme (e.g. facility-based, area-based, or distance-based), (2) the degree to which tolls vary over time, (3) other dimensions of toll differentiation, and (4) technology. The review by A. de Palma and R. Lindsey (Appendix A) explaining these dimensions is focused on technology for the congestion charges. The main technology categories are roadside-only systems employing digital photography, tag and beacon systems that use short-range microwave technology, and in vehicle-only systems based on either satellite or cellular network communications. Road pricing technology should be chosen to best meet objectives. In addition to congestion relief, road pricing can be used to internalize the costs of emissions, accidents, noise, and road damage. It can also be used to pay for parking, to generate revenues, and to implement the beneficiary principle that the costs of roads should be paid by those who use them. Pricing congestion efficiently is arguably the most demanding goal in terms of differentiation by vehicle characteristics, location, time of day, and real-time driving conditions. This suggests that congestion pricing should drive the technology choice. But the economics of congestion pricing are more attractive if the technology that is chosen can perform other functions.

The authors conclude that the low-tech options Automated Number Plate Recognition (ANPR) and Dedicated Short Range Communications (DSRC) are better suited for tolling individual facilities and urban areas where congestion is severe. The high-tech option (Satellite and Cellular technologies) appear to be more economical for pricing at larger geographical scales. While simplicity favours the low-tech option, interoperability requires the high-tech option. Incompatibility of congestion charging systems may become a problem as tolling becomes more widespread.

Given the many arguments for and against low-tech and high-tech options, and the importance of ancillary concerns such as public acceptability, it appears that both low-tech and high-tech options will be pursued in the near term. Given their advantages in terms of scale economies, value-added services, and revenue generation as a supplement or replacement for fuel taxes, it seems plausible that either Satellite or Cellular technologies will come into widespread use in the longer term. If so, it makes sense to use them for congestion pricing as well as other functions.

2.2. Using a Conventional Four Stage Model to Model Congestion Charging

In many countries it is standard to use a conventional four stage model in the decision process of major transportation projects, such as investment in new roads. These models have been developed over a long period of time and are often rich in geographical data. However, during the last years it has become clear that, while appropriate for modelling of infrastructure investments, the four stage model show major deficits when modelling the effects of demand management tools, such as congestion charging.

The four stage model, described in detail for example by Ortuzar and Willumsen (1994), includes the stages: trip generation, trip distribution, mode split and assignment (Figure 1). Trip generation takes as input land use data and population characteristics and results in number of trips produced and attracted in each zone. The second stage represents choice of destination and each produced trip is assigned a destination zone, resulting in an origin-destination matrix of trips. This matrix is split into one matrix with demand for each mode in the third stage and assigned to the corresponding networks (usually car and public transport networks) in the last stage. Algers (2000) note that the model is not always applied stage after stage, rather it is sometimes applied as a simultaneous hierarchical model. Since the later choices may affect previous choices, the four stage model should be iterated until convergence is reached. For example, a congested road network effects mode choice such that more people choose public transport, but as more travellers choose public transport the road network is not as congested anymore, and so on. The feed-back loop is a result of criticism of the four stage model by Bonsall (1997), but is still disregarded in some cases.

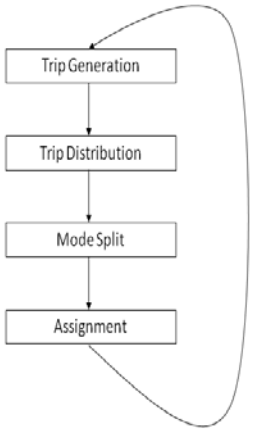


Figure 1: The four stages in a conventional transport model

The conventional model for transport planning does not have to look exactly as the four stage model described above, actually many variants exists. Algers (Algers 2000) summarizes the main features common to conventional models:

- Zonal basis
- Demand in trips or tours
- Static structure
- Structured set of travel choices
- Iteration to equilibrium

Several problems arise when applying a conventional four stage model for design and prediction of congestion charging schemes:

First, the four stage model typically has a static structure in which there is no time dimension (as was also stated as one of the main features above) and is run separately for peak hour demand levels and inter-peak demand levels. A peak hour factor, representing the percentage of trips that occur during

peak hour, is then usually used to weigh together the two solutions. If the time-dimension is present in a transport model for an urban area, it is almost always a time-of-day choice between broad time periods such as peak and inter-peak, not changes within the peak period (see e.g. (RAND Europe 2004) for an example of time-of-day choice between broad time periods). Peak-hour factors may be a reasonable simplification for some applications of transport models. However in networks that suffer from congestion more advanced modelling of temporal aspects are advisable since peak-hour factors are independent of congestion levels. A model that uses peak-hour factors will tend to overestimate peak hour congestion levels in future years: demand increases but is in the model not allowed to spread outside the peak hour, because the percent of trips simulated during peak hour is fixed. If simulation of car trips is part of a larger modelling system the result may also be an overrating of number of users switching to public transport, due to the very long peak hour car travel times. Also, this kind of model may overrate the benefit of adding capacity in a congested area, since temporal effects that result in car-users switching back to the peak hour are not accounted for.

Second, also related to the static structure of the model, is that delay on links is calculated using volume-delay functions. This representation of delay is not able to capture the effect of capacity limitation, such as queue spillback with blocking back of upstream links, or capacity reductions due to signals and conflicting flows at intersections.

Third, congestion charging in urban areas is typically applied in or around the city centre, which has the potential to encourage park-and-ride. However, in the conventional four stage model there is a structured set of travel choices, splitting up of demand matrix into one matrix for each mode, and separate assignment of car and public transport trips, which makes it difficult to model trips that contain more than one mode, such as park-and-ride trips.

Fourth, the four stage model typically uses one average value of time for each trip purpose. Values of time vary however within each trip purpose, which is very important for prediction of congestion charging effects, since trips are sorted, based on their value of time.

Modelling frameworks for prediction of congestion charging effects are reviewed in (Milne 2009) and (Koh & Shepherd 2006). They show that several urban areas use conventional four stage models for prediction of congestion charging effects. This has for example been the case in Copenhagen (Jovicic & Hansen 2003), Hong Kong (Wheway & Cheuk 1999), London (ROCOL 2000), Oslo (Larsen & Ostmoe 2001), the Netherlands (Fox et al. 2003), Singapore (Le & Lim 2003) and Stockholm (Transek 2003), in all of which a strategic transport model coupled with static assignment (many of them using the static assignment package EMME/2) was used for design and prediction of congestion charging schemes. Some of the models listed above included modifications to deal with congestion charging such as time-of-day choice (Oslo, Hong Kong and London) and improved modelling of delays at junctions (Singapore).

Note that the three cities were congestion charging has actually been implemented, Singapore, London and Stockholm, all used some form of conventional four stage model for design and prediction of their charging schemes. Below is a description of the models used in these three cities and a discussion about the predictability of the models.

2.3. Singapore – The LTA Strategic Transport Model

The LTA (Land Transport Authority) strategic transport model includes three time periods (AM peak hour, PM peak hour and off-peak) and five trip purposes (home-based work, home-based-shopping, home-based-leisure, home-based-business, non-home-based) (Le & Lim 2003). Peak hour factors by trip purpose, mode and area are used to represent percent of trips that occur during peak hours. The choice dimensions available are: trip generation, trip distribution, mode choice and route choice, i.e. the typical steps of a four stage model. The static assignment model EMME/2 is used to calculate route choice and resulting travel times for each origin-destination pair. Iteration of the model is

performed but the first stage, trip generation, is not included in iterations, i.e. the generation of trips is calculated once and for all from land use data and population characteristics.

Some modification of the LTA strategic transport model has been done to improve prediction of charging schemes. This modification concerns the modelling of delay at intersections. In the standard LTA model link delays were calculated based on speed-flow curves and traffic assigned to the network based on this information. The improved model is called the *iterative approach* (Le & Lim 2003). In this approach, additional intersection coding is performed were the user adds data related to if the intersection is signalized or not, green time, cycle time etc. A special delay function for signalized movement is determined, which has an upper bound on the delay for a given link. Thus the capacity restriction of the link is accounted for.

No data could be found on the ability of the LTA model to make correct predictions of congestion charging schemes.

2.4. London – The APRIL/AREAL Model

One of the first models that tried to forecast the effects of congestion charging in a large urban network was APRIL. In the so called ROCOL study (Review Of Charging Options for London), that preceded implementation of congestion charging in London, the APRIL model was developed further to be able to forecast effects of area charging schemes and was renamed to AREAL (ROCOL 2000). The difference between AREAL and APRIL is that in AREAL there is a model for the share of households that will buy an area licence coupon, depending on charged amount, household income and number of cars in the household.

AREAL includes seven time periods (06.00-07.00, 07.00-10.00, 10.00-11.00, 11.00-15.00, 15.00-16.00, 16.00-19.00, 19.00-20.00) and eight trip purposes (home to/from work, home to/from employer's business, home to/from education, home to/from other, non-home-based to/from employer's business, non-home-based to/from other, light goods vehicles, heavy goods vehicles). The choice dimensions available are: trip generation, trip distribution, mode choice, time period choice and route choice, but not all choice dimensions are available for all trip purposes. For example only home to/from other trips were allowed to change destination. The full table showing which choice dimensions are applicable to which trip purposes can be found in Annex E in (ROCOL 2000). Four of the trip purposes were modelled as tours (home to/from work, home to/from employer's business, home to/from education, home to/from other) and the other four as trips (non-home-based to/from employer's business, non-home-based to/from other, light goods vehicles, heavy goods vehicles).

Even though AREAL included some nice features for modelling of area charging, such as tour modelling and modelling of propensity to buy licence coupons, it turned out that the model could not make reliable forecasts of changes in speeds and corresponding travel times that would result from introducing congestion charging. The main problem was in representation of delay, i.e. the speed-flow relationships did not produce improvements in speeds consistent with observations. Another model called LTS, which had a more detailed road network, was also applied to London but showed the same deficits in speed forecasts. The problems were so large that no definite conclusions could be drawn for impacts of charging in different parts of London (ROCOL 2000).

2.5. Stockholm – The SAMPERS Model

The Swedish travel demand forecasting tool SAMPERS is described in (Beser & Algers 2001). In design and prediction of congestion charges SAMPERS' regional model for Stockholm was applied, which is called SAM and covers Stockholm County. Previously a study had been made with the four stage model FREDRIK (Mattsson 2003), which is similar to SAMPERS' regional model for Stockholm, but not integrated in a national modelling system and estimated on an older travel behaviour survey.

SAMPERS' regional models include five trip purposes (home-based-work, business, school, social, recreation, other). Number of trips is estimated for six modes (car as driver, car as passenger, bus, commuter train, bicycle and walk). The model structure is a hierarchical nested logit model for the travel choices trip/no trip, mode choice and destination choice, which work on a tour basis. EMME/2 is used as assignment model for both car and public transport. To determine the amount of trips that occur during peak hour, a peak hour factor for each trip purpose is applied.

A comparison between modelled and observed effects of congestion charging in Stockholm is given by Engelson and van Amelsfort (2011). This comparison shows that flow decrease is overestimated by SAMPERS compared to observations: 7.30-9.00 am the traffic flow change is -13% in observations, but -29% in SAMPERS when weekdays before the Stockholm Trial are compared to weekdays during the Trial. The change in speed is however underestimated: average speed increases with 13% in observations, but only with 5,3% in SAMPERS. This shows the inability of static models to correctly predict changes in speed (or equivalently travel time) as a result of a flow decrease.

3. Problems with static models

Because of inherently dynamic nature of congestion it is desirable to use a model that takes into account the temporal phenomena like queue accumulation and discharge, blocking back, and interaction of vehicle flows at intersections. However, as listed by Koh and Shepherd (2007), most models used for design of road user charging use static (Oslo, Copenhagen, the Netherlands, Singapore) or semi-dynamic (London) traffic assignment. For design of congestion charging system in Stockholm the model Sampers (Beser Hugosson and Algers, 2002) based on static traffic assignment was used, and for *a posteriori* cost benefit analysis results from the same model were supplemented by traffic counts and travel time measurements (Eliasson, 2009a).

The reason of using the static assignment for the task is twofold. First, the static assignment with separable volume-delay functions has nice mathematical properties of solution existence and uniqueness which is important for scenario comparison, and there exist convergent algorithms (Patriksson, 1994) implemented in commercial software. Second, the required extensive data collection, development and calibration of a dynamic model for the whole city is often considered a too time consuming and expensive effort by planners and decision makers. On the contrary, a conventional transportation model involving aggregate travel demand modeling and static traffic assignment with separable volume-delay functions may already exist for the study area or can be developed by established methods within a relatively short time period and used for design and evaluation of congestion charges.

However, there are several pitfalls in using static models for effects of congestion charges. Paulley (2000) points out that in most cases, even when the validation of the model is acceptable, it does not necessary imply that delays are adequately represented because the response of the speed flow curve to delay has to do with the gradient rather than the absolute speed levels. Li (2002) has shown that knowledge of the speed-flow relationship is critical for determination of the optimal charges while absence of the information about demand-cost relationship can be compensated by sequential trials of change levels. However, the static volume delay functions cannot describe relationship between travel demand and travel time on a level detailed enough for road charging application. As noted by Bonsall et al (2005), attempting to model the fact that the delay by the vehicles that arrive in a queue early on is greater than that created by the vehicles that arrive towards the end of the life of the queue requires a dynamic simulation network. Lo and Szeto (2005) have shown by small network examples that, in the presence of queue spillback, the static paradigm cannot adequately portray the congestion phenomenon. Boyles et al (2005) have demonstrated that a static model considerably underestimates congestion levels compared to a dynamic one for the same scenario.

Although the static models may be quite suitable for strategic land use and transportation planning they lack the temporal dimension and therefore may produce results that are too far from the actual

outcome of the introduction of charging system especially in terms of travel time savings and social benefit.

Engelson and van Amelsfort (2011), see Appendix B, compare forecast by the model Sampers used for the design of congestion charge system in Stockholm with observations before and during the Stockholm trial. The comparison has shown that the drawbacks of static assignment with separable costs are crucial for forecast of effect of congestion charges. The changes in traffic volumes are overestimated and the changes in travel times are underestimated and this problem cannot be fixed by reasonable modification of VDF. Even 10 times steeper VDF will not provide the appropriate level of changes. In order to produce a reasonably correct forecast of changes in travel conditions with charges it is essential for the analysis to use a model that recognizes intersection interactions and blocking back. Engelson and van Amelsfort (2011) conclude that dynamic assignment is crucial for informed decision on introduction of measures that aim at relieving congestion.

4. Important issues for the modelling of congestion charges

The static model remains a basic tool for the mathematical description of congested networks. The static model does, however, omit important features of congestion. The static model is hence unsatisfactory for a number of purposes. The main feature that the static model omits is that congestion varies over the day, with pronounced AM and PM peaks in most cities. Travel times can easily increase by a factor two from the beginning to the height of the peak. To design and evaluate policies for tackling congestion it is necessary to recognize these variations. There are a number of fundamental features of congested demand peaks that a model should take into account.

First, travelers choose not only a route, but also a departure time in response to how congestion varies over the course of a day. When a policy is implemented that affects peak congestion, travellers respond by changing departure time.

Second, travelers incur more than just monetary costs and travel time costs when they make a trip. Travellers have preferences regarding the timing of trips and deviations from the preferred timing are costly. Such scheduling costs are comparable in magnitude to congestion-delay costs as a fraction of total user costs. These scheduling costs are by nature ignored in static models. This means that static models cannot reveal the effect of policies that affect scheduling costs.

Third, many relevant policies can only be described within a dynamic model. A congestion toll or parking fee that varies over time as congestion increases and decreases is an obvious example.

The basic dynamic model, the **bottleneck model** introduced by Vickrey (1969), starts directly from the above observations regarding within-day dynamics. It is therefore well suited to analyze policies that rely on these dynamics. The bottleneck represents any road segment that constitutes a binding capacity constraint. The bottleneck allows users to pass only at some fixed rate. An area of economic literature has grown out of this concept, exploring a number of issues in the context of the basic bottleneck model, e.g. equilibrium, social optimum, decentralization of the social optimum via pricing, second best pricing (including step tolls), elastic demand, heterogeneous individuals, small networks (routes in parallel and routes in series), stochastic capacity and demand, alternative treatments of congestion, and pricing on large networks. These issues are discussed by de Palma and Fosgerau (Appendix C). The basic model has also been extended to include mode choice, parking congestion, modeling of the evening commute and non-commuting trips. The research stream initiated with M. Ben-Akiva had more focus on numerical computation, and it has led, amongst other development, to the METROPOLIS software for large networks.

Dynamic models can be used to study a variety of policies that cannot be studied with static models. These include road pricing with a time-varying component, flexible work hours, staggered work hours, dynamic access control, and ramp metering used to differentiate capacity allocation. Pricing policies are much more effective when tolls depend on the time of the day, for stylised as well as for real networks.

The current state of the topic of dynamic congestion modelling provides a range of general insights from small stylised models. Numerical simulation models exist to deal with the complexities of real size networks. In between, there is a large gap. Numerical simulation has the drawback that it must rely on particular assumptions, which may or may not provide good approximations to the object of interest. So a main motivation for continued theoretical research into dynamic models of congestion is the desire for increased generality. The fewer assumptions required for a conclusion, the more certain we can be that it applies.

Acknowledgement

The research was financed by Sweden, France, Denmark, Finland and Switzerland within ERA-NET TRANSPORT under the theme SURPRICE ("Road User Charging for Passenger Vehicles").

We would like to thank all those who have accompanied us in the development of this project. Firstly we want to thank Prof. Mogens Fosgerau, Technical University of Denmark and Prof. Robin Lindsey, University of British Columbia (Canada) as co-authors of theoretical contributions. Then, we would like to thank DRAST, especially Gérard Brun. Two persons helped us in every aspects of the project. Kiarash Motamedi, and Fabrice Marchal. Without their help and suggestions this project won't be finished on time. A special thanks to all the participants of PREDIT for their comments and suggestion and finally, our gratitude to the Technical University of Berlin for organizing a project meeting for the participants of Silverpolis.

References :

Algers, S., (2000), Appendix A: New look at multi-modal modelling: From a mainstream point of view. *In: DS Consultancy (ed): A new look at multi-modal modelling*, Report to the DETR.

Arnott, R., A. de Palma and R. Lindsey, (1994), The welfare effects of congestion tolls with heterogeneous commuters, *Journal of Transport Economics and Policy*, 28(2), 139–161.

Beser, M. and S., Algers, (2001), SAMPERS—The new Swedish national travel demand forecasting tool, *National Transport Models: Recent Developments and Prospects*, 101–118.

Bonsall, P.W., (1997), Principles of transport analysis and forecasting, Chapter 5 in C A O'Flaherty (ed): *Transportation Planning and Traffic Engineering*, Arnold: London.

Bonsall, P., H. Link, K.Toepel, A. Ricci, R.Enei, J.C. Martin, C. Román, A. Voltes, H. Meersman, T.Pauwels, E. Van deVoorde and T.Vanelslander, (2005), Information requirements for analysis of optimal complexity, *Deliverable 2 of GRACE (Generalisation of Research on Accounts and Cost Estimation)*, Funded by Sixth Framework Programme. ITS, University of Leeds, Leeds.

Boyles, S., S. V. Ukkusuri, S. T. Waller and K. M. Kockelman, (2006), A comparison of static and dynamic traffic assignment under tolls in the Dallas-Fort Worth region. Presented at the TRB Annual Meeting.

de Palma, A. and R. Lindsey, (2006), Modelling and evaluation of road pricing in Paris, *Transport Policy*, 13(2), 115–126.

de Palma, A. and F. Marchal, (2002), Real cases applications of the fully dynamic METROPOLIS tool-box: an advocacy for large-scale mesoscopic transportation systems, *Networks and Spatial Economics*, 2(4), 347–369.

de Palma, A., M. Kilani and R. Lindsey, (2005), Congestion pricing on a road network: A study using the dynamic equilibrium simulator METROPOLIS, *Transportation Research Part A*, 39(7-9), 588–611.

de Palma, A., F. Marchal and Y. Nesterov, (1997), METROPOLIS: Modular system for dynamic traffic simulation, *Transportation Research Record*, 1607, 178–184.

Eliasson, J. and L.G. Mattsson, (2006), Equity effects of congestion pricing: Quantitative methodology and a case study for Stockholm, *Transportation Research Part A*, 40(7), 602–620.

Eliasson, J., (2009a), A cost-benefit analysis of the Stockholm congestion charging system, *Transportation Research Part A: Policy and Practice*, 43, 468-480.

Engelson, L. and D. van Amelsfort, (2011), The role of volume-delay functions in forecast and evaluation of congestion charging schemes, application to Stockholm, In *Proceedings of the Kuhmo Nectar Conference*, Stockholm, June 2011.

Fox, J., Daly, A. and H. Gunn, (2003), Review of RAND Europe's transport demand model systems, Rand Europe, prepared for TRL Limited, Available at: http://www.its.leeds.ac.uk/projects/distillate/outputs/Deliverable%20F%20Appendix%20A.pdf [Accessed May 30, 2011].

Glazer, A. and E. Niskanen, 2000, Which consumers benefit from congestion tolls?, *Journal of Transport Economics and Policy*, 34(1), 43–53.

Jovicic, G. and C.O. Hansen, (2003), A passenger travel demand model for Copenhagen, *Transportation Research Part A: Policy and Practice*, 37(4), 333–349.

Koh, A. and S. Shepherd, (2006), Issues in the modelling of road user charging, Distillate Project F, Appendix A, ITS Leeds, Available at: http://www.its.leeds.ac.uk/projects/distillate/outputs/Deliverable%20F%20Appendix%20A.pdf [Accessed May 30, 2011].

Kristoffersson, I., (2011), Impacts of time-varying cordon pricing: Validation and application of mesoscopic model for Stockholm, *Transport Policy*, In Press, Available online 6 August 2011, DOI: 10.1016/j.tranpol.2011.06.006.

Kristoffersson, I. and L. Engelson, (2009), A dynamic transportation model for the Stockholm area: Implementation issues regarding departure time choice and OD-pair reduction, *Networks and Spatial Economics*, 9(4), 551–573.

Kristoffersson, I. and L. Engelson, (2011), Alternative road pricing schemes and their equity effects: Results of simulations for Stockholm, In *Proceedings of the TRB 90th Annual Meeting*, Washington, D.C., January 2011.

Larsen, O.I. and K. Ostmoe, (2001), The experience of urban toll cordons in Norway: lessons for the future, *Journal of Transport Economics and Policy (JTEP)*, 35(3), p.457–471.

Le, H. and W.L. Lim, (2003), Junction modelling in a strategic transport model, In *17th International EMME/2 Users' Conference*.

Li, Micjael Z. F., (2002), The role of speed-flow relationship in congestion pricing implementation with an application to Singapore, *Transportation Research Part B*, 36, 731-754.

Lo, H.K. and W.Y. Szeto, (2005), Road pricing modeling for hyper-congestion, *Transportation Research Part A: Policy and Practice*, 39, 705-722.

Mattsson, L.G., (2003), Modelling road pricing reform in Stockholm, Unpublished paper, Royal Institute of Technology, Department of Infrastructure, p.1–26.

Milne, D.S., (2009), CURACAO - Coordination of urban road user charging organisatonal issues, Chapter 6 in Deliverable D2: State of the Art Review, Institute for Transport Studies, University of Leeds, Leeds, United Kingdom.

Nielsen, O.A., A. Daly and R. Frederiksen, (2002), A stochastic route choice model for car travellers in the Copenhagen region, *Networks and Spatial Economics*, 2(4), 327–346.

Odeck, J., J. Rekdal and T. Hamre, (2003), The socio-economic benefits of moving from cordon toll to congestion pricing: The case of Oslo, In *Proceedings of the TRB 82nd Annual Meeting*, Washington, D.C., January 2003.

Ortuzar, J.D. and L.G. Willumsen, (1994), *Modelling transport*, Wiley, New York.

Patriksson, M., (1994), The Traffic Assignment Problem: Models and Methods, V.S.P. Intl Science.

Paulley, N., (2000), Advice on Modelling of Congestion charging or Tolling options for Multimodal Studies, London DfT.

Pigou, A.C., (1920), The Economics of Welfare, 4th. London: Macnillam.

RAND Europe, (2004), PRISM West Midlands Time-of-day choice models, Available at: http://217.206.77.227/prism/Downloads/task/Task%203.pdf [Accessed May 23, 2011].

Rich, J. and O.A. Nielsen, (2007), A socio-economic assessment of proposed road user charging schemes in Copenhagen, *Transport Policy*, 14(4), 330–345.

ROCOL, (2000), Road charging options for London, a technical assessment, Annex E - Behvioural modelling and model parameters, London: HMSO. Available at: http://webarchive.nationalarchives.gov.uk/20100528142817/http://www.gos.gov.uk/497417/docs/204399/rocol_ch8_annexes_121k.pdf [Accessed May 26, 2011].

Small, K., (1983), The incidence of congestion tolls on urban highways, Journal of urban economics 13(1), 90–111.

Transek, (2003), Försök med miljöavgifter i Stockholm, Underlag för utformning och genomförandeplan, Consultant report to Stockholm City. (In Swedish), Available at: http://www.stockholmsforsoket.se/upload/MAK/Bakgrundsdokument/MAKunderlag.pdf [Accessed May 25, 2011].

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

Wheway, J. and K. Cheuk, (1999), Implementation of road pricing model with the EMME/2 macro language, In 1st Asian EMME/2 Users Conference Shanghai.

Annex A: Dynamic and Static Congestion Models

Reference:

de Palma, A. and M. Fosgerau (2011), Dynamic and Static Congestion Models: a Review. In *Handbook in Transport Economics*, A. de Palma, R. Lindsey, E. Quinet et R. Vickerman, (eds.), Edgar Elgard.

Abstract

We begin by providing an overview of the conventional static equilibrium approach. In such model both the flow of trips and congestion delay are assumed to be constant. A drawback of the static model is that the time interval during which travel occurs is not specified so that the model cannot describe changes in the duration of congestion that result from changes in demand or capacity. This limitation is overcome in the Vickrey/Arnott, de Palma Lindsey bottleneck model, which combines congestion in the form of queuing behind a bottleneck with users' trip-timing preferences and departure time decisions. We derive the user equilibrium and social optimum for the basic bottleneck model, and explain how the optimum can be decentralized using a time-varying toll. They then review some extensions of the basic model that encompass elastic demand, user heterogeneity, stochastic demand and capacity and small networks. We conclude by identifying some unresolved modelling issues that apply not only to the bottleneck model but to trip-timing preferences and congestion dynamics in general.

1. Introduction[1]

This Paper provides a brief introduction to dynamic congestion models, based on the Vickrey (1969) bottleneck model which has become the main workhorse model for economic analysis of situations involving congestion dynamics.

The word *dynamic* can have several possible meanings. One possibility is that it relates to the way traffic systems evolve and users learn from day to day. In the context of the bottleneck model, it relates to intra-day timing, *i.e.* to the interdependencies between traffic congestion at different times within a given day.

We shall discuss dynamic approaches against the background of static models. Static models assume that congestion is constant over some given time period. A congestion law provides the travel time as a function of the entering flow. The time dimension is not explicitly involved: all quantities are computed as single figures specific to a time period.

The basic static model considers a network comprising nodes and links. The nodes are centroids of zones, associating trip ends within a zone with a point that is a node in the network. Links connect the nodes. A cost function describes the cost of using each link. Congestion means that the cost increases as the number of users of the link increases. The demand is given by the origin-destination (O-D) matrix, indicating the number of trips between pairs of nodes. The solution involves the choice of route within the network for each O-D pair. Traffic volume on each link, the travel cost of using each link, the cost of making each trip, and the total travel cost for all users all depend on these route-choice decisions.

Each user for each O-D pair is assumed to choose a route in the network that minimises the sum of link costs for the trip. But users compete for the same space and the route choices of users in one O-D pair affect the costs experienced by other users through congestion. We can imagine a process where users keep revising their route choices in response to the route choices of other users. We seek an equilibrium in which no user can reduce his cost by choosing a different route. This equilibrium concept is due to Wardrop (1952). This problem was first given a mathematical formulation and solution for a general network by Beckmann, McGuire and Winston (1956). We will discuss the static model in more detail below in the context of simple networks.

The static model remains a basic tool for the mathematical description of congested networks. The static model does, however, omit important features of congestion. The static model is hence unsatisfactory for a number of purposes.

The main feature that the static model omits is that congestion varies over the day, with pronounced AM and PM peaks in most cities. Travel times can easily increase by a factor two from the beginning to the height of the peak. To design and evaluate policies for tackling congestion it is necessary to recognize these variations. There are a number of fundamental features of congested demand peaks that a model should take into account.

First, travelers choose not only a route, but also a departure time in response to how congestion varies over the course of a day. When a policy is implemented that affects peak congestion, travellers respond by changing departure time. The departure time changes are systematic on average and can be observed in the aggregate temporal shape of the peak. Think, for example, of car traffic entering the central business district (CBD) of some large city. The number of travellers reaching their workplaces per hour is fixed at the capacity rate during the morning peak. So if the number of workplaces in the CBD increases, the duration of the morning peak must increase too. Similarly, if capacity is increased the duration of the peak will shrink. The duration of the peak thus depends on

[1]We would like to thank Robin Lindsey for many useful comments, suggestions and references.

both demand and capacity. Such observations suggest that trip timing is endogenous and speak in favour of dynamic models.

Second, travelers incur more than just monetary costs and travel time costs when they make a trip. Travellers have preferences regarding the timing of trips and deviations from the preferred timing are costly. Such scheduling costs are comparable in magnitude to congestion-delay costs as a fraction of total user costs. These scheduling costs are by nature ignored in static models. This means that static models cannot reveal the effect of policies that affect scheduling costs.

Third, many relevant policies can only be described within a dynamic model. A congestion toll or parking fee that varies over time as congestion increases and decreases is an obvious example.

The basic dynamic model discussed in this chapter, the bottleneck model, starts directly from the above observations regarding within-day dynamics. It is therefore well suited to analyse policies that rely on these dynamics. It was introduced by William Vickrey (1969). Arnott, de Palma, and Lindsey (1993a) revisited and extended this seminal but almost forgotten model. It is a tractable model and it leads to a number of important insights. The model features one origin-destination pair (let us say residence and workplace), one route, and one bottleneck. The bottleneck represents any road segment that constitutes a binding capacity constraint. The bottleneck allows users to pass only at some fixed rate. There is a continuum of users and it takes some positive interval of time for them all to pass the bottleneck. Users are identical and they wish to arrive at the destination at the same ideal time t^*. Because of the bottleneck, all but one user must arrive either before or after t^*. Deviation from t^* represents a cost for users. They also incur a travel time cost, which includes free flow travel time and delays in the bottleneck. Individuals choose a departure time to minimise the sum of schedule delay and travel time costs.

To analyze this situation, we consider an equilibrium in which no traveller has incentive to change his departure time choice. This is an instance of a Nash equilibrium (Haurie&Marcotte, 1985), which is the natural generalization of Wardrop equilibrium. Individuals are identical and therefore they experience the same cost in equilibrium. One might wonder whether the Nash equilibrium concept has any counterpart in the real world. We see Nash equilibrium as a benchmark. Like anything else in our models, it is an idealisation, describing a situation that we hope is not too far from reality. The appeal of Nash equilibrium is that it is a rest point for any dynamic mechanism whereby informed travellers revise their (departure time) choice, if they do not achieve the maximum utility available to them.

Travelers incur the same generalised travel cost in equilibrium, but they have different trips. Some depart early, experience only a short delay at the bottleneck, but arrive early at work. Others avoid queuing delay by departing late, but arrive also late at work. Those who arrive near the preferred arrival time will experience most congestion and have the longest travel time. In this way, the bottleneck model describes a congested demand peak with a queue that first builds up and then dissolves.

The endogenous choice of the departure time was independently studied by de Palma, Ben-Akiva, Lefèvre&Litinas (1983), who proposed a dynamic model incorporating a random utility departure time choice model and a generalized queuing model. In contrast to the Vickrey bottleneck model, where the capacity constraint is either active or not, the supply model of de Palma *et al.* shifts smoothly from the uncongested to the congested regime.

An area of economic literature has grown out of these two initial contributions, exploring a number of issues in the context of the basic bottleneck model: *e.g.*, equilibrium, social optimum, decentralization of the social optimum via pricing, second best pricing (including step tolls), elastic demand, heterogeneous individuals, small networks (routes in parallel and routes in series), stochastic capacity and demand, alternative treatments of congestion, and pricing on large networks. The basic model has also been extended to include mode choice, parking congestion, modelling of

the evening commute and non-commuting trips. The research stream initiated with M. Ben-Akiva had more focus on numerical computation, and it has led, amongst other development, to the METROPOLIS software for large networks, discussed below.

This chapter first reviews the simple static model of congestion, where time is not explicitly considered. This serves as a background for the dynamic model. We then introduce the basic bottleneck model and continue to discuss some of the extensions mentioned.

2. The static model of congestion [2]

2.1. Static networks

We begin with a simple example. Consider a fixed number $N > 0$ of travellers having two routes available. The travellers split with $n_1 > 0$ on the first route and $n_2 > 0$ on the second route, where $n_1 + n_2 = N$. The cost associated with each route is taken to be a linear function of traffic such that the average cost on route i is $C_i(n_i) = a_i + b_i n_i$. The cost is a so-called generalised cost, combining monetary cost and travel time in a single monetary equivalent. The Nash equilibrium occurs when no traveller wants to change route, which requires that $C_1(n_1) = C_2(n_2)$. Solving this equation leads to the equilibrium solution[3]

$$n_1^e = \frac{a_2 - a_1}{b_1 + b_2} + \frac{b_2}{b_1 + b_2} N, n_2^e = N - n_1^e .$$

The Nash equilibrium has every traveller minimise his/her own cost. We can alternatively consider social optimum where the total cost for all travellers is minimised. In general the social optimum is not a Nash equilibrium. The social optimum minimises the total cost function

$$\min_{n_1, n_2} W(n_1, n_2) = n_1 C_1(n_1) + n_2 C_2(n_2) .$$

The total cost associated with use of route i is $n_i C_i(n_i)$. The marginal cost of an additional user is

$$\frac{d\left[n_i C_i(n_i)\right]}{dn_i} = C_i(n_i) + b_i n_i .$$

In this expression, $C_i(n_i)$ is the cost paid by the marginal user. The remainder $b_i n_i$ is an externality: it is the part of the increase in the total cost that is not borne by the additional user. The first-order condition for social optimum requires equal marginal costs, or

$$C_1(n_1) + b_1 n_1 = C_2(n_2) + b_2 n_2 . \quad (1)$$

The only difference between this and the first-order condition for the equilibrium is the terms representing the externalities of the two routes. The externalities are zero if $b_i = 0$, i=1,2, *i.e.* if adding an additional user does not lead to increased travel cost. In this case, the social optimum would be the same as the equilibrium.

---

[2] For a more detailed analysis of congestion in the static model see the chapter by Santos and Verhoef in Handbook in Transport Economics,Volume 1 & 2 (de Palma, A., R. Lindsey, E. Quinet& R. Vickerman 2011), Edgar Elgard, sous presse.

[3] We assume the parameters are such that this equation leads to positive flows on each route.

The social optimum has

$$n_1^o = \frac{a_2 - a_1}{2b_1 + 2b_2} + \frac{2b_2}{2b_1 + 2b_2} N, n_2^0 = N - n_1^o . \qquad (2)$$

The solution is written in this way to emphasize the similarity to the Nash equilibrium. The only difference between the optimum and the equilibrium outcomes is that the marginal costs, the b_i, have been replaced by $2b_i$ in the expression for the optimum outcome. This indicates that the optimum can be achieved as an equilibrium outcome by setting a toll equal to $n_i b_i$ on each of the two routes. This has the effect of doubling the variable cost from the perspective of users and the expression in (2) then becomes the equilibrium outcome.

2.2. Elastic demand

The discussion so far has considered a fixed number of travellers N. We now allow demand to be elastic, limiting attention to just one route. Travellers on this route are identical, except for different willingness to pay to travel. Figure 1 shows a downward-sloping inverse demand curve $D(N)$ to reflect that demand decreases as the cost increases. The curve $C(N)$ is again an average cost curve expressing the cost that each traveller incurs. The curve $MC(N)$ is a marginal cost curve, expressing the marginal change in total cost following a marginal increase in the number of travellers; in other words[4]

$$MC(N) = C(N) + N \cdot C'(N).$$

When the cost curve is increasing, the marginal cost curve will lie above the cost curve.

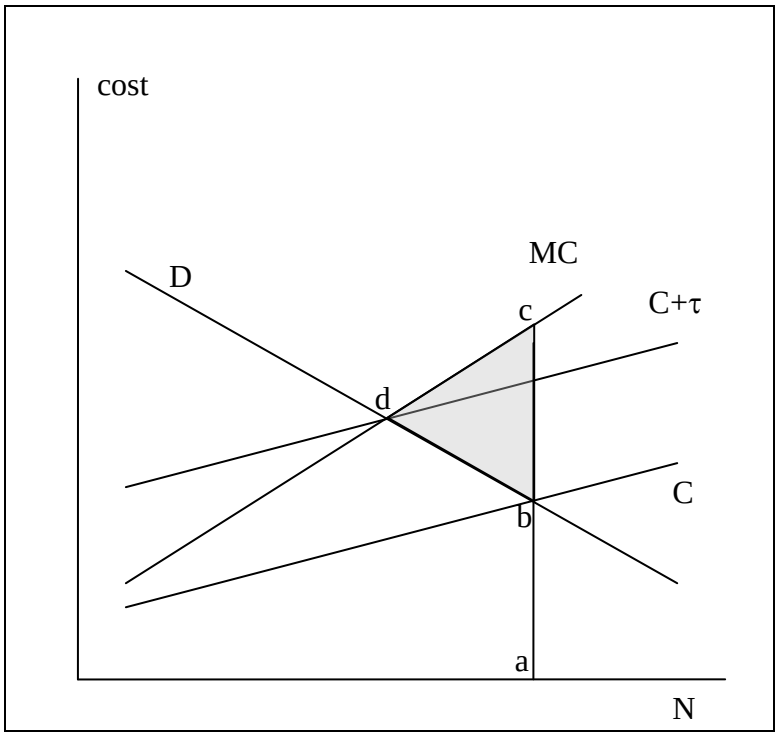


Figure 1 A static model

[4] C' denotes the derivative of C.

The equilibrium occurs at the intersection of the demand curve with the average cost curve at the point b. The marginal traveller at this point is indifferent between travelling and not travelling, he faces a cost corresponding to the line segment a-b and a benefit of the same size. For travellers in aggregate, however, the cost of adding the marginal traveller is given by the MC curve. For the marginal traveller at point b, this cost corresponds to the line segment a-c. So the last traveller imposes a net loss corresponding to the line segment b-c on the group of all travellers. If usage was reduced to the point where the MC curve crosses the demand curve, then the corresponding loss is zero for the traveller at the point d. The total loss in market equilibrium is then represented by the shaded triangle b-c-d on the figure.

The optimal toll, labeled τ in Figure 1, implements the optimum at the point d, where the private benefit is equal to the marginal cost. The toll is required because drivers ignore the costs they impose on other drivers. The toll is just the difference, evaluated at the social optimum, between the marginal cost and the average cost, *i.e.* the externality.

3. The basic bottleneck model

We now introduce the basic Vickrey bottleneck model in its simplest form. Consider a continuum of $N > 0$ identical travellers, who all make a trip. They have to pass a bottleneck, which is located d_1 time units from the trip origin and d_2 time units from the destination. Denote the time of arrival at the bottleneck of a traveller by t and the exit time from the bottleneck as a. The situation is illustrated in Figure 2. A traveller departs from the origin at time $t - d_1$ and arrives at the bottleneck at time t. There he/she is delayed until time $a \geq t$ at which time he/she exits from the bottleneck to arrive at the destination at time $a + d_2$.

Each traveler has a scheduling cost expressing his/her preferences concerning the timing of the trip. Travelers are assumed to have a preferred arrival time t^* and they dislike arriving earlier or later at the destination. Travelers also prefer the trip to be as quick as possible. For a trip that starts at time t_1 and ends at time t_2, consider then a cost of the form

$$c(t_1, t_2) = \alpha \cdot (t_2 - t_1) + \beta \cdot \max(t^* - t_2, 0) + \gamma \cdot \max(t_2 - t^*, 0), \quad (3)$$

where $0 < \beta, 0 < \gamma$ and $\beta < \alpha$. In this formulation, α is the marginal cost of travel time, β is the marginal cost of arriving earlier than the preferred arrival time, γ is the marginal cost of arriving later, and these values are constant. The deviation $t_2 - t^*$ between the actual arrival time and the preferred arrival time is called schedule delay and it is possible to speak of schedule delay early and schedule delay late, depending on the sign of the schedule delay.[5]

[5] Small (1982) tested a range of formulations of scheduling preferences, including the $\alpha - \beta - \gamma$ preferences as a special case.

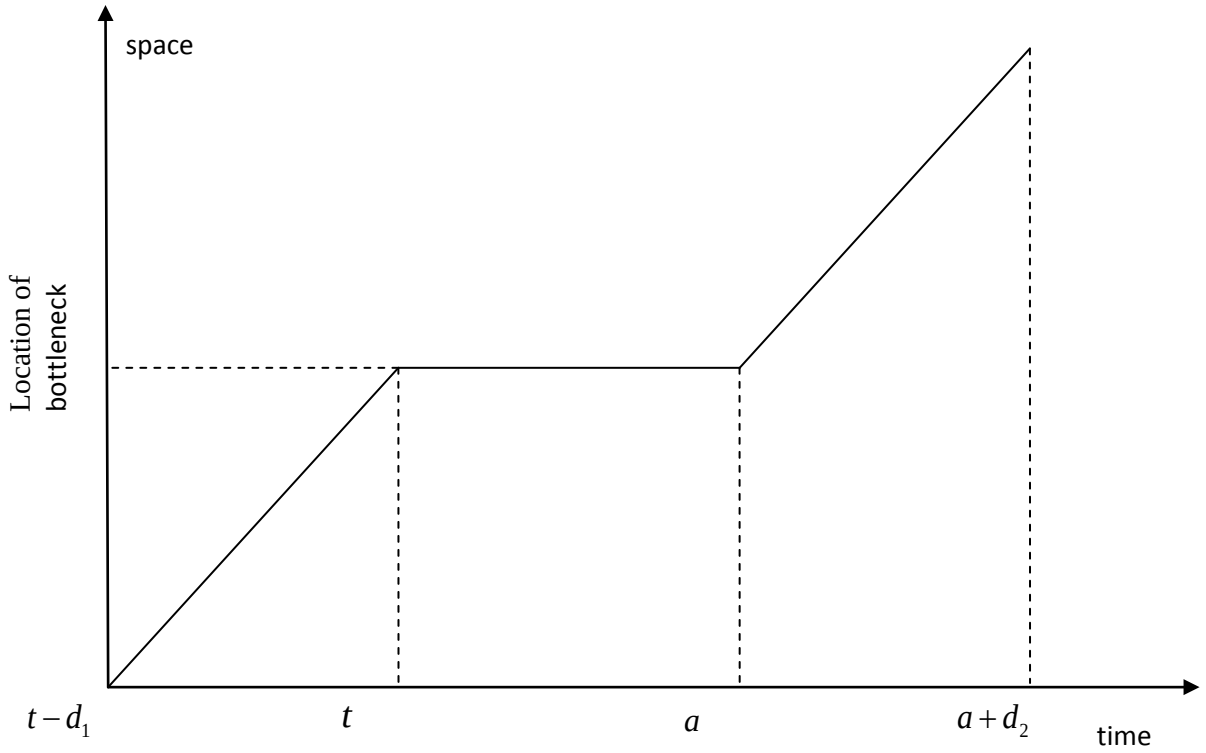


Figure 2 Trip timing

This cost formulation has become colloquially known as $\alpha-\beta-\gamma$ preferences. Later, we shall consider scheduling cost of a general form.

The travel time d_1 between the origin and the bottleneck adds the same constant amount to the scheduling cost of all travellers and so it can be set to zero without affecting the behaviour of travellers in the model. Similarly, the travel time d_2 between the bottleneck and the destination can be set to zero by redefining the preferred arrival time. So without loss of generality we may let $d_1 = d_2 = 0$. This means that the time of departure is the same as the time of arrival at the bottleneck and that the time of exit from the bottleneck is the same as the time of arrival at the destination.

Travelers depart from the origin according to an aggregate schedule, described in terms of the cumulative departure rate R, where $R(a)$ is the number of travellers who have departed before time a. So R is similar to a cumulative distribution function: it is proportional to the probability that a random traveller has departed before time a. R is increasing, since travellers never return. Moreover, $R(-\infty)=0$ and $R(\infty)=N$. The departure rate $\rho(a)=R'(a)$, wherever R is differentiable.

The bottleneck can serve at most s travellers per time unit. Travellers who have not yet been served wait before the bottleneck. The bottleneck serves travellers in the sequence in which they arrived (first-in-first-out or FIFO). The bottleneck capacity is always used if there are travellers waiting before it.

Recall that Nash equilibrium is defined as a situation in which no traveller is able to decrease his cost by choosing a different departure time. Since travellers are identical, this definition reduces to the

requirement that all travellers experience the same cost and that the cost would be higher for departure times that are not chosen by any travellers.

Denote the interval of departures and arrivals as $I = [a_0, a_1]$. Let us consider some properties of Nash equilibrium. First, there will be queue from the time the first traveller departs until the last traveller departs, since otherwise there would be a gap in the queue and somebody could move into the gap to decrease cost. Second, the queue will end at the time the last traveller departs, since otherwise he/she could wait until the queue was gone and reduce cost. This shows that the departure interval is just long enough for all travellers to pass the bottleneck. Third, as the cost of the first and the last travellers are equal and since they experience no queue, they must experience the same cost due to schedule delay. These insights are summarised in the followingequations.

$$a_1 - a_0 = N / s, \qquad (4)$$

$$\beta \cdot (t^* - a_0) = \gamma \cdot (a_1 - t^*). \qquad (5)$$

Equation (4) ensures that arrivals take place during an interval that is just long enough that all travelers can pass the bottleneck. Equation (5) ensures that no traveler will want to depart at any time outside I.

Solving these two equations leads to

$$a_0 = t^* - \frac{\gamma}{\beta + \gamma} \frac{N}{s},$$

$$a_1 = t^* + \frac{\beta}{\beta + \gamma} \frac{N}{s}$$

and the equilibrium cost for every traveler is

$$\frac{\beta\gamma}{\beta + \gamma} \frac{N}{s} \equiv \delta \frac{N}{s}.$$

This is linear in the number of travelers and so the simple static model could be viewed as a reduced form of the dynamic model.

Equations (4) and (5) are extremely useful in that they determine the equilibrium cost of travelers as a function of the number of travelers and the bottleneck capacity. The total cost is then $\delta N^2 / s$ with corresponding marginal cost $2\delta N / s$, of which half is internal cost to each traveller and the other half is external. The marginal change in total cost following a change in capacity s is $-\delta N^2 / s^2$. Since there is no toll, price equal travel cost: $p^e = \delta N / s,$ that is price is a function of *N* and *s*. The function is this a reduced-form supply function, which is very usefull, especially in analytical work, together with a trip demand function.

There is always a queue during the interval I. This means that the bottleneck capacity is fully utilised and hence that sd travellers pass the bottleneck during an interval of length d. At time a, a total of $R(a)$ travellers have entered the bottleneck, taking a total time of $R(a)/s$ to pass. The first traveller enters and exits the bottleneck at time a_0. Hence a traveller arriving at bottleneck at time a exits at time $a_0 + R(a)/s$. Travellers are identical so they incur the same scheduling cost in equilibrium. Normalising $t^* = 0$, itemergesthat

$$\delta \frac{N}{s} = \alpha \cdot \frac{R(a)}{s} + \beta \max\left(-a_0 - \frac{R(a)}{s}, 0\right) + \gamma \max\left(a_0 + \frac{R(a)}{s}, 0\right).$$

Differentiating this expression leads to

$$\rho(a) = \begin{cases} s\frac{\alpha}{\alpha - \beta}, a_0 + \frac{R(a)}{s} \leq 0 \\ s\frac{\alpha}{\alpha + \gamma}, a_0 + \frac{R(a)}{s} > 0 \end{cases}$$

during interval I. A few observations are immediately available. Initially the departure rate is constant and higher than s (since $\beta < \alpha$). It is high until the traveller who arrives exactly on time. Later travellers depart at a constant rate which is lower than s.

Figure 3 shows the resulting departure schedule. The horizontal axis is time and the vertical axis is the number of departures, ranging from 0 to N. The thick kinked curve is the cumulative departure rate R. Departures begin at time a_0 and end at time a_1 with $R(a_1) = N$. The line segment connecting point a_0 to point e represents the number of travellers served by the bottleneck, it has slope s.

The first departures take place at a rate larger than capacity and queue builds up. For example, at time a, the number of travellers who have departed corresponds to the length of the segment $a - c$, while the number of travellers who have been served by the bottleneck corresponds to the length of the segment $a - b$. Thus the queue at that time has length corresponding to the segment $b - c$. The travellers in the queue at time a will all have been served by time d, which is then the time at which the traveller departing at time a is served by the bottleneck. The time spent in the bottleneck equals the length of the queue at the time of departure divided by the capacity.

The traveler departing at time d exits the bottleneck exactly at time a_*. Therefore the departure rate drops below capacity at this time and the queue begins to dissolve. It also follows that the queue reaches its maximum length at time d.

For the top half of the figure, the horizontal time axis refers both to the departure from the origin and to the arrival time at the destination. For the bottom half of the figure, the time axis instead refers to the arrival time at the destination. The shaded areas on the bottom half of Figure 3 show the composition of the scheduling cost throughout the peak. The first traveller arrives early and is not delayed in the bottleneck so his cost is $\beta \cdot (a_* - a_0)$. Later travellers do not arrive as early, but are delayed more in the queue and incur the same trip cost. The traveller who arrives at the preferred arrival time is the most delayed and his trip cost comprises solely travel time cost. Later arrivals are less delayed in the queue, but arrive later at the destination. The last traveller is not delayed in the bottleneck at all, but arrives last at the destination and incurs a cost of $\gamma \cdot (a_1 - a_*)$.

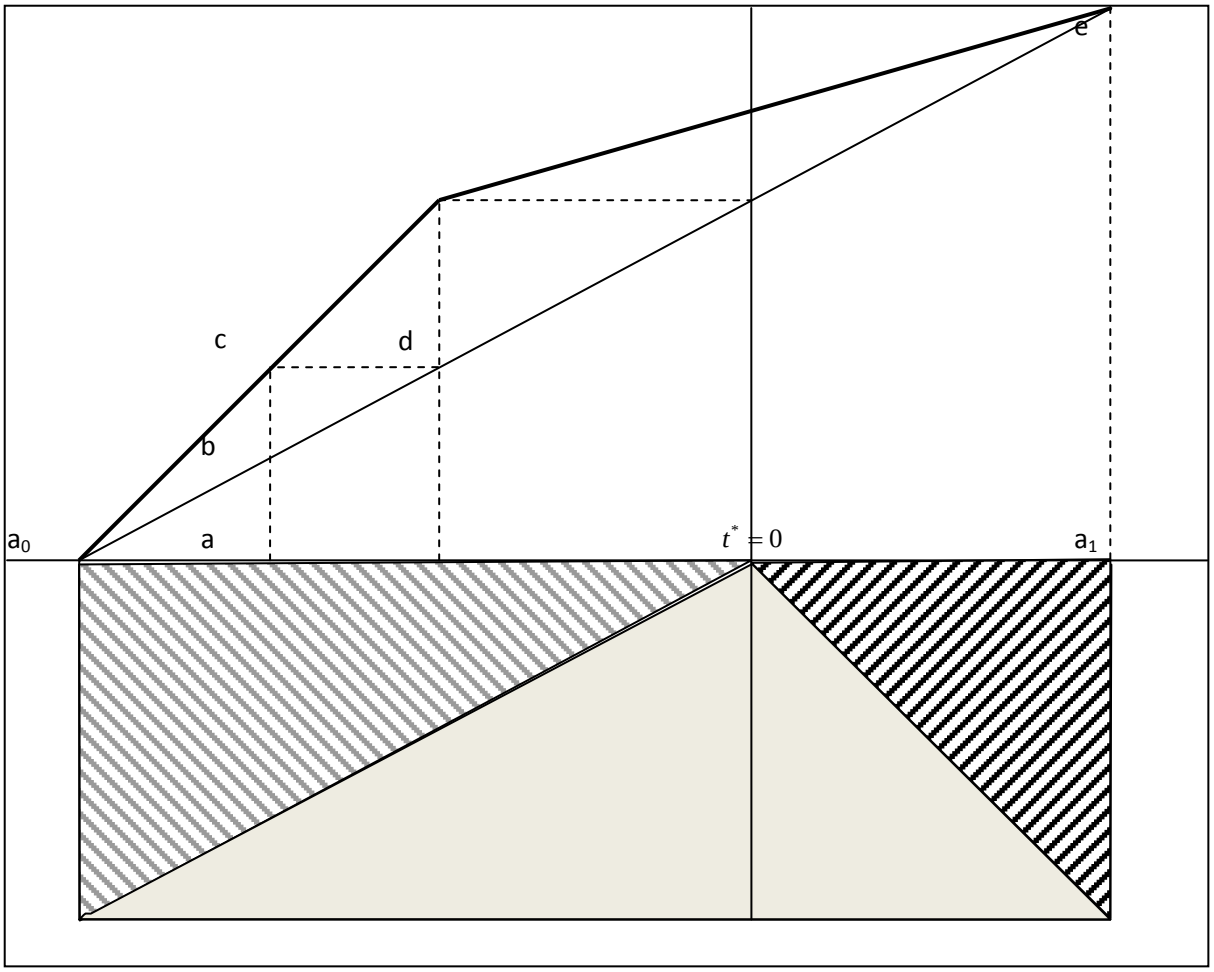


Figure 3 Equilibrium departure schedule under $\alpha - \beta - \gamma$ preferences

3.1. Optimal tolling

The queue that arises in equilibrium in the bottleneck model is sheer waste. It generates no benefit at all. If travellers could be induced to depart at the capacity rate s during the equilibrium interval I, then there would be no queue. All travellers (except the very first and the very last) would gain from reduced travel time while arriving at the destination at exactly the same time as in equilibrium. A main insight of the bottleneck model is that it is possible to achieve this outcome through the application of a toll.

So consider a time varying toll $\tau(\cdot) \geq 0$ charged at the time of arrival at the bottleneck. We make the additional behavioural assumption that travellers choose departure time to minimise the sum of the toll and the trip cost. We restrict attention to tolls that have $\tau(a_0) = \tau(a_1) = 0$ and are zero outside the departure interval I. This means that (4) and (5) still apply. If the toll is well-behaved, in ways to be explained below, then Nash equilibrium exists and departures still occur in the interval I. Therefore the equilibrium cost is the same as in the no-toll equilibrium discussed above.

Travellers do not lose, but somebody else may gain since revenue from the toll can be used for other purposes. The size of the toll revenue is

$$\int_{a_0}^{a_1} \tau(s) ds, \qquad (6)$$

and this represents a net welfare gain.

Since the cost must be constant in equilibrium, we have

$$\tau(a) = \delta \frac{N}{s} - c\left(a, \frac{R(a)}{s} + a_0\right), \qquad (7)$$

where R is now the departure rate that results when the toll is imposed. It is (intuitively) clear that maximal efficiency is attained when the toll revenue is as large as it can be without destroying the equilibrium. Increasing $\tau(a)$ in (7) will reduce $R(a)$.[6] Moreover, the queue cannot be negative and so we must require that $R(a) \geq s(a - a_0)$. Therefore the maximal toll maintains zero queue and the least possible cumulative departure rate, *i.e.* $R(a) = s \cdot (a - a_0)$. This corresponds to a constant departure rate $\rho(a) = s$. The optimal toll is

$$\tau(a) = \delta \frac{N}{s} - c(a, a) = \delta \frac{N}{s} - \beta \cdot \max(-a, 0) - \gamma \cdot \max(a, 0)$$

for $a \in I$ and zero otherwise. This toll is initially zero at time a_0. Then it increases at the rate β until it reaches a maximum of $\delta N / s$ at time 0. It then decreases at the rate γ until it is again zero at time a_1. The optimal toll corresponds to the grey shaded area in Figure 3. In a sense, it just replaces the cost of queueing by a toll. The efficiency gain is achieved because queueing is pure waste whereas the toll revenue is just a transfer.

3.2. Elastic demand

The discussion of the bottleneck model so far has assumed demand to be inelastic. A natural extension is to assume that the number of travelers deciding to participate in the peak depends on the equilibrium cost (Arnott*et al.*, 1993a). The trip cost

$$p = \tau(a) + c(a, a_0 + R(a)/s) \quad (8)$$

is the same for all travelers in equilibrium. This implies that the total toll payment is $N \cdot (p - \overline{c})$, where $\overline{c}$ is the average scheduling cost of travellers. Let $N(\cdot) > 0, N'(\cdot) < 0$ be a downward sloping demand function such that $N(p)$ is the realised demand.

This is a very convenient way to extend the model: conditional on any equilibrium number of travellers, the properties of equilibrium are exactly the same as in the inelastic case. The equilibrium number of travellers is uniquely determined since demand is decreasing as a function of the equilibrium cost of travellers while the equilibrium cost of travellers is increasing as a function of the number of travellers. This simplicity comes, however, at a cost as it requires separability between trip timing on the one hand and participation on the other.

The separability of trip timing and participation implies that the optimal toll with elastic demand is the same as in the case of inelastic demand. To see this, note first that the optimal toll is able to remove queuing, so the average cost of travellers remains equal to $\delta N / s$. Consider the following welfare function

[6] Since $c_2(t_1, t_2) < 0$. This follows since $\beta < \alpha$. We use subscripts to denote partial derivatives.

$$W(p) = \int_p^{\infty} N(s)\,ds + N \cdot (p - \overline{c}),$$

i.e. the sum of consumer surplus and the total toll revenue. To find the welfare optimising toll, note that

$$\overline{c} = \frac{s}{N} \int_{a_0}^{a_1} c(a,a)\,da,$$

which can be shown to imply that

$$\frac{\partial \overline{c}}{\partial p} = \frac{N'(p)}{N(p)} \left(\delta N / s - \overline{c} \right).$$

Using this to evaluate the first-order condition for maximum of $W(p)$ leads to $p = \delta N / s$. That is, the optimal price should equal the equilibrium scheduling cost. Using (8) shows that the optimal toll is $\tau(a) = \delta N / s - c(a,a)$, which is the same as in the case of inelastic demand.

3.3. Optimal capacity and self-financing

Consider now a situation in which the optimal toll applies while capacity s is supplied at cost $K(s) \geq 0$, with $K' > 0$. We extend the social welfare function with the cost of capacity provision

$$W(p,s) = \int_p^{\infty} N(r)\,dr + N \cdot (p - \overline{c}) - K(s).$$

For any given capacity s, the optimal value of $\tau(a) = \delta N / s - c(a,a)$ is as shown above. Note that

$$\frac{\partial \overline{c}}{\partial s} = \frac{1}{s} \left(\delta N / s - \overline{c} \right).$$

This can be used to show that capacity is optimal when $sK'(s) = N \cdot (p - \overline{c})$. That is, the revenue from the optimal toll is equal to $sK'(s)$.

This finding leads directly to the self-financing theorem for the bottleneck model. If capacity is produced at constant returns to scale, *i.e.* if $K(s) = sK'(s)$ with $K'(s)$ constant, then the optimal toll exactly finances the optimal capacity $K(s) = N \cdot (p - \overline{c})$. If there are increasing returns to scale, then $K(s) > sK'(s)$, in which case the optimal toll cannot finance the optimal capacity.

The self-financing result is also called the cost recovery theorem. It is an instance of a general self-financing theorem by Mohring&Harwitz (1962), which assumes that travel cost is homogenous of degree zero in capacity and use. A number of results on self-financing are summarized by Verhoef&Mohring (2009).

The optimal capacity can be computed in the three regimes: no toll, coarse step toll and optimal fine toll. It can be shown that the optimal capacity is the lowest for the optimal fine toll, intermediary for the coarse toll and larger for the no toll regime (see ADL, 1993, for a proof). Note that these proofs are correct with inelastic (and elastic) demand.

4. Scheduling preferences

4.1. General formulation

The $\alpha-\beta-\gamma$ formulation of scheduling cost used above is a special case of more general scheduling preferences, introduced in this section. Below we revisit the bottleneck model from the perspective of these general scheduling preferences.

In order to describe the traveller choice of trip timing in a more general way, we formulate scheduling preferences for a given trip in the form of scheduling utility $u(t_1,t_2)$, where t_1 is the departure time and t_2 is the arrival time,. We shall make minimal assumptions regarding the specification of u.

It is natural to require that $u_1 = du/dt_1 > 0$, such that it is always preferred to depart later, given t_2.[7] Similarly, requiring $u_2 = du/dt_2 < 0$ ensures that arriving earlier is always preferred, given t_1. A marginal increase in travel time then always leads to a utility loss, since travellers will either have to depart earlier or arrive later. Define the function $v(a) = u(a,a)$ as the scheduling utility that a traveller would receive if travel was instantaneous. Assume that v is quasi-concave and attains maximum at $v(t^*)$. This assures that for any $d > 0$ there is a unique solution to the equation $v(a) = v(a+d)$. It also implies that v is increasing for $a < t^*$ and decreasing for $a > t^*$.

We incorporate monetary cost by considering utility to be $u-\tau$.In some cases it is more convenient to talk about cost, which will then be the negative of utility, *i.e.* $\tau-u$. In either case, it is implied that there is separability between scheduling and monetary cost. That is, a constant cost does not affect the preferences regarding trip timing.

In some situations it is necessary to specify scheduling utility further by imposing a certain functional form. For example, the $\alpha-\beta-\gamma$ formulation specifies the scheduling cost completely up to a few parameters. Such restriction can be necessary for reasons of identification in econometric work, but in general it is preferable to specify as little as possible, since restricting the model entails the risk of introducing errors. In theoretical models it is similarly preferable to work with general formulations, since otherwise there is a risk that the results one may obtain depend on the specific formulation.

In some cases it may be considered acceptable to impose a separability condition, just as we have done in the case of monetary cost and trip timing. The timing of the trip is given by a departure time and an arrival time and we work under the assumption that these times are all that matter about trip timing. The travel time is the difference between the departure time and the arrival time. We could equivalently describe trip timing in terms of travel time and arrival time or in terms of travel time and departure time. From the perspective of general scheduling utility $u(t_1,t_2)$, this leads to three possibilities for introducing a separability condition.

$$u(t_1,t_2) = f(t_2-t_2) + g(t_1)$$
$$u(t_1,t_2) = f(t_2-t_1) + g(t_2)$$
$$u(t_1,t_2) = f(t_1) + g(t_2)$$

The first condition would say that scheduling utility is separable in travel time and departure time. The second condition would say instead that scheduling utility is separable in travel time and arrival time. The $\alpha-\beta-\gamma$ scheduling cost is a special case of this second possibility: Changing the travel

[7] We use subscripts to denote partial derivatives.

time does not affect the traveller preferences regarding arrival time and vice versa. The third possible separability condition is used in the Vickrey (1973) formulation of scheduling preferences that we will consider in the next section. Here scheduling utility is separable in departure time and arrival time. That is, changing departure time, does not affect the preferences regarding arrival time and vice versa.

The concept of the preferred arrival time t^* was used to define the $\alpha-\beta-\gamma$ scheduling cost. It makes sense to talk about a preferred arrival time when there is separability in travel time and arrival time, since then the preferred arrival time is not affected by the travel time. Without this separability, there is no single preferred arrival time since the preferred time to arrive depends on the travel time. If instead scheduling utility is separable in departure time and travel time, then we would want to talk about a preferred departure time. In some contexts, for example the PM commute from work to home, this might be a more natural concept. In general, neither the concept of a preferred arrival time nor a preferred departure time may be relevant. We shall now discuss Vickrey (1973) scheduling preferences, which are separable in departure time and arrival time.

4.2. Vickrey (1973) scheduling preferences

Consider an individual travelling between two locations indexed by $i=1,2$. He derives utility at the time dependent rate η_i at location i. Let us say he starts the day at time T_1 at location 1 and ends the day at time T_2 at location 2. If he departs from location 1 at time t_1 and arrives (later) at location 2 at time t_2, then he obtains scheduling utility

$$u(t_1,t_2)=\int_{T_1}^{t_1}\eta_1(s)\,ds+\int_{t_2}^{T_2}\eta_2(s)\,ds. \qquad (9)$$

The formulation is illustrated in Figure 4.

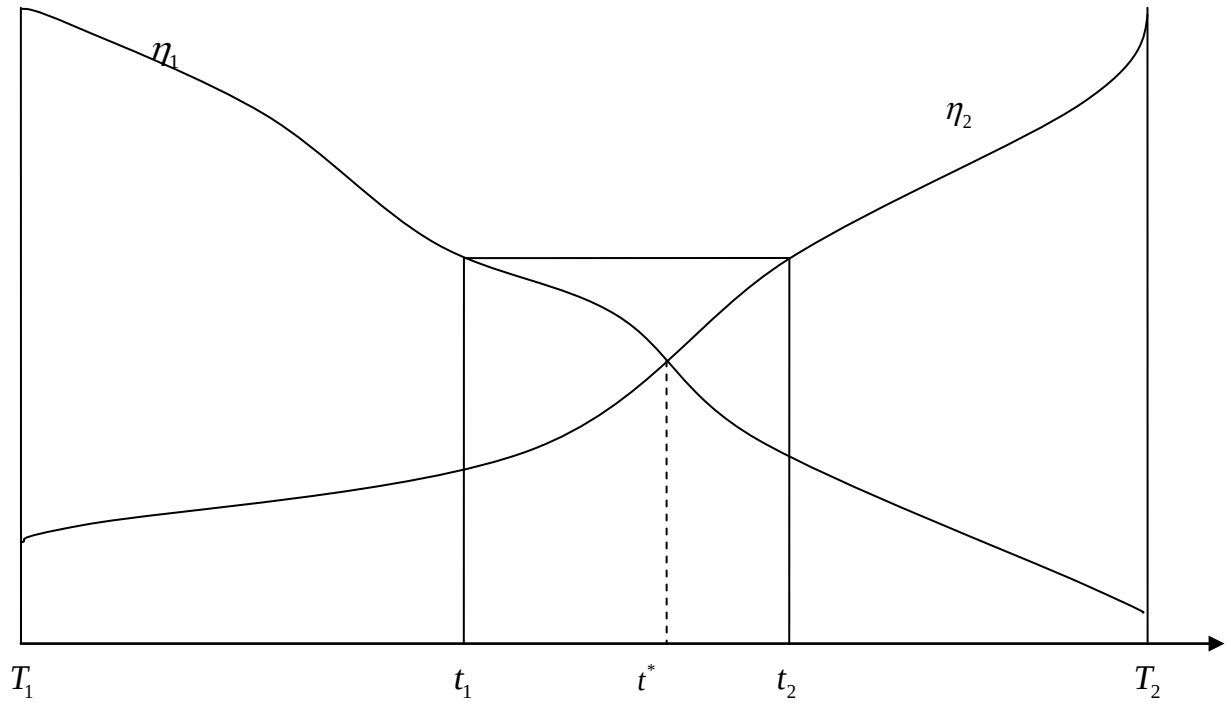


Figure 4 Vickrey (1973) scheduling preferences

Note that when T_1 and T_2 are fixed, these numbers can be replaced by arbitrary numbers in equation(9) without affecting the implied preferences. Assume that $\eta_1 > 0,\ \eta'_1 < 0$, $\eta_2 > 0,\ \eta'_2 < 0$ and that there is a point in time, t^*, where $\eta_1\left(t^*\right) = \eta_2\left(t^*\right)$. Speaking in terms of the morning commute these conditions imply that a traveller prefers to be at home or at work to travelling, that his/her marginal utility of staying later at home is decreasing, that his/her marginal utility of arriving earlier at work is also decreasing, and that there is a time (t^*) when he/she would optimally transfer from home to work if instant travel was possible. Given a travel time of d , he/she would optimally depart at the time $t(d)$ depending on d when $\eta_1\left(t(d)\right) = \eta_2\left(t(d)+d\right)$. It is straightforward to derive that his/her value of time would be

$$-\frac{\partial u\left(t(d), t(d)+d\right)}{\partial d} = \eta_2\left(t(d)+d\right).$$

This is strictly increasing as a function of d . Using survey data on stated choice, Tseng &Verhoef (2008) provide empirical estimates of time varying utility rates corresponding to the Vickrey (1973) model.

4.3. The cost of travel time variability

When travel time is random and travellers are risk averse, the random travel time variability leads to additional cost, the cost of travel time variability. Both Vickrey formulations of scheduling preferences are useful for deriving measures of the cost of travel time variability as well as of the scheduling impact of the headway of scheduled services. Such cost measures can be useful to incorporate elements of dynamic congestion in reduced form in static models. Consider a traveller who is about to undertake a given trip. The travel time for the trip is random from the perspective of the traveller. While he/she does not know the travel time outcome before making the trip, the traveller knows the travel time distribution. The travel time distribution is independent of the departure time of the traveller. The latter is a strong assumption but necessary for the results

The traveler is assumed to choose his departure time optimally, so as to maximise his/her expected scheduling utility. That makes the expected scheduling utility a function just of the travel time distribution. Therefore it is possible in principle to evaluate how the expected scheduling utility depends on the travel time distribution. Simple expressions are available for the two Vickrey specifications of scheduling preferences.

In the case of $\alpha - \beta - \gamma$ preferences, Fosgerau&Karlstrom (2010) show that the expected trip cost with optimal departure time is

$$\alpha \cdot \mu + \sigma \cdot (\beta + \gamma) \int_{\gamma/(\beta+\gamma)}^{1} \Phi^{-1}(s)\, ds,$$

which is linear in the mean and in the standard deviation of travel time. This is a practical advantage in applications. The expression depends on the shape of the travel time distribution through the presence of Φ in the integral and so Φ must be taken into account if the marginal value of standard deviation of travel time is to be transferred from one setting to another. In the same vein, Fosgerau (2009) uses $\alpha - \beta - \gamma$ scheduling cost to derive simple expressions for the value of headway for scheduled services. In the case of Vickrey (1973) scheduling preferences with linear utility rates, Fosgerau&Engelson (2010) carry out a parallel exercise. They show that with random travel time and unconstrained choice of departure time, the expected scheduling cost with the optimal choice of departure time is linear in travel time, travel time squared and the variance of

travel time. Parallel results are also provided for the value of headway for scheduled services. In contrast to the case of $\alpha - \beta - \gamma$ scheduling cost, it is possible also to derive a simple expression for the expected scheduling cost for the case of a scheduled service with random travel time.

4.4. The bottleneck model revisited

The results discussed above for the basic bottleneck model survive in some form with more general scheduling preferences. The setup of the model is as before, the only change is that now travellers are only assumed to have scheduling preferences of the general form discussed above. Without loss of generality we may again consider $d_1 = d_2 = 0$, since the exact form of scheduling preferences is not specified.

It is easy to argue, using the same argument as in the simple case, that Nash equilibrium requires departures in an interval $I = [a_0, a_1]$ satisfying

$$a_1 - a_0 = N / s, \qquad (10)$$

$$v(a_0) = v(a_1). \qquad (11)$$

This is illustrated in Figure 5. Moreover, the queue has length zero at time a_0 and a_1 but it is strictly positive at any time in the interior of this interval. The second condition (11) has a unique solution since v is quasiconcave and it ensures that no traveler will want to depart at any time outside I.

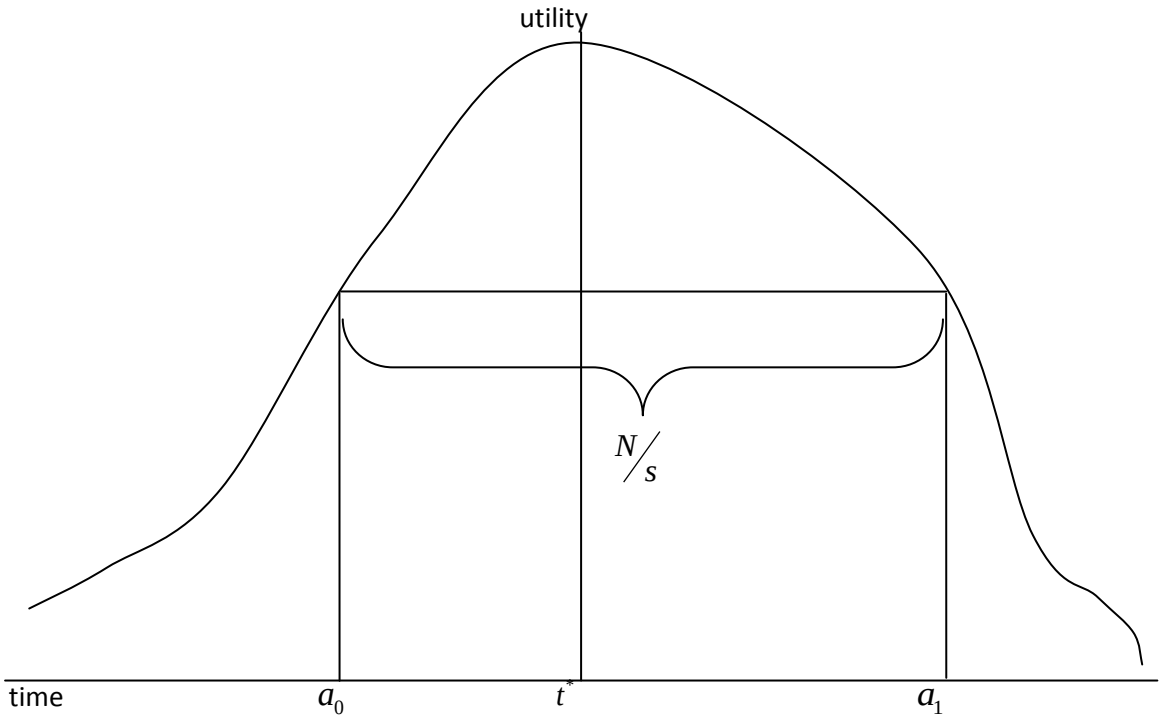


Figure 5 The function v and the equilibrium departure interval

Equations (10) and (11) determine the equilibrium utility of travelers as a function of the number of travelers and the bottleneck capacity. It is then straightforward to derive the marginal external congestion cost and the marginal benefit of capacity expansion.

As in the basic model, there is always a queue during the interval I and a traveller arriving at the bottleneck at time a exits at time $a_0 + R(a)/s$. Travellers are identical so they achieve the same scheduling utility in equilibrium

$$v(a_0) = u\left(a, a_0 + \frac{R(a)}{s}\right).$$

Consider now a time varying toll $\tau(\cdot) \geq 0$ charged at the time of arrival at the bottleneck. We restrict attention to tolls that have $\tau(a_0) = \tau(a_1) = 0$ and are zero outside the departure interval I. This means that equations (10) and (11) still apply. If the toll is not too large, then Nash equilibrium exists with departures still in the interval I.[8]Therefore the equilibrium utility $v(a_0)$ is the same as in the no-toll equilibrium. As in the basic model, the optimal toll maintains the departure rate at capacity. The optimal toll is then given by $\tau(a) = v(a) - v(a_0)$ for $a \in I$ and zero otherwise.

The conclusions regarding elastic demand extend to the case of general scheduling preferences. That is, the optimal toll is still $p = \tau - u$, which is the same as in the case of inelastic demand. The conclusions regarding optimal capacity and self-financing also carry over to the general case. That is, if capacity is supplied at constant cost and optimally chosen, then the optimal toll exactly finances the capacity cost.

5. Extensions of the bottleneck model

The bottleneck model is useful in many ways. It generates a number of insights concerning dynamic congestion, while being still relatively simple and tractable. The model is useful if the mechanisms it describes are representative of the real world. It is, however, a highly stylised description of actual congested networks. It is therefore of interest to extend the model by introducing more relevant features. Such an exercise has two main purposes. One is to gauge the robustness of the conclusions of the basic model. We can have greater confidence in conclusions that survive in more general versions of the model. The other main purpose is to generate new insights that were not available with the basic model. This section proceeds with a presentation of some of the extensions of the bottleneck model available in the literature.

5.1. Second best pricing

The optimal toll described above varies continuously over time. A real toll could do the same to any relevant degree of precision, but there remains the problem that travellers may not be able to understand such a complex pricing structure. Moreover, there may be technological reasons for varying tolls less frequently. Acceptability of road pricing is also a fundamental issue (see, *e.g.* de Palma, Lindsey &Proost, 2007, on this issue).

Such considerations have led researchers to consider tolls that vary in steps. In the context of the bottleneck, ADL (1990) consider the simplest step toll, namely a toll that is positive and constant during some interval and zero otherwise. Such a toll has also been called a coarse toll.

The discrete jumps of such a toll generate some new properties of the resulting equilibrium. Three groups of travellers can be identified according to whether they travel before, during or after the tolling period. Figure 6 compares the cumulative departure curve in the step-toll equilibrium and compares it with the no-toll equilibrium. Consider first the time before the toll is turned on. The cost of the last traveller not to pay the toll should be the same as the cost of the first traveller to pay the

[8]Provided that the toll does not decrease too quickly. A quickly decreasing toll may induce travelers to avoid certain departure times, which leads to unused capacity.

toll. To achieve this equality, there must be a period with no departures between these two travellers. Early in the morning travellers depart at a high rate, they pay no toll and consequently depart at the same rate as they would in no-toll equilibrium. Just before the departure time at which travellers would begin to pay the toll, departures cease for a while and the queue dissipates gradually as travellers are served by the bottleneck.

Departures start again when the queue has diminished just enough for the toll payment to be compensated by lower queueing time. The optimal single step toll is timed such that the queue has just disappeared at the time the toll kicks in. Departures for the group of travellers paying the toll then continue following the pattern analysed above. The toll is constant for these travellers and hence does not affect the departure rates. The departure rate is consequently high until the time at which a traveller arrives at the destination exactly on time, and then it drops to a lower level. The optimal single step toll is timed such that the queue has just disappeared at the time the toll lifts.

A new phenomenon emerges relating to the third and final group of travellers who do not pay the toll. As shown in Figure 6, there is no queue at the moment before they depart. But the first traveller to depart must have the same cost in equilibrium as the other travellers. This can only happen if there is a mass departure at this time. In a mass departure, travellers depart so closely together that their sequence in the queue is random. In this case, travellers are assumed to account for their expected trip cost.

It turns out that all the remaining travellers depart at once under the optimal coarse toll if $\alpha < \gamma$ as has been found in most empirical studies. On average they are better off than a traveller who waits until the queue has gone before departing. But all travellers must achieve the same expected cost in equilibrium. Therefore the first traveller departs later under the coarse toll than under no toll.

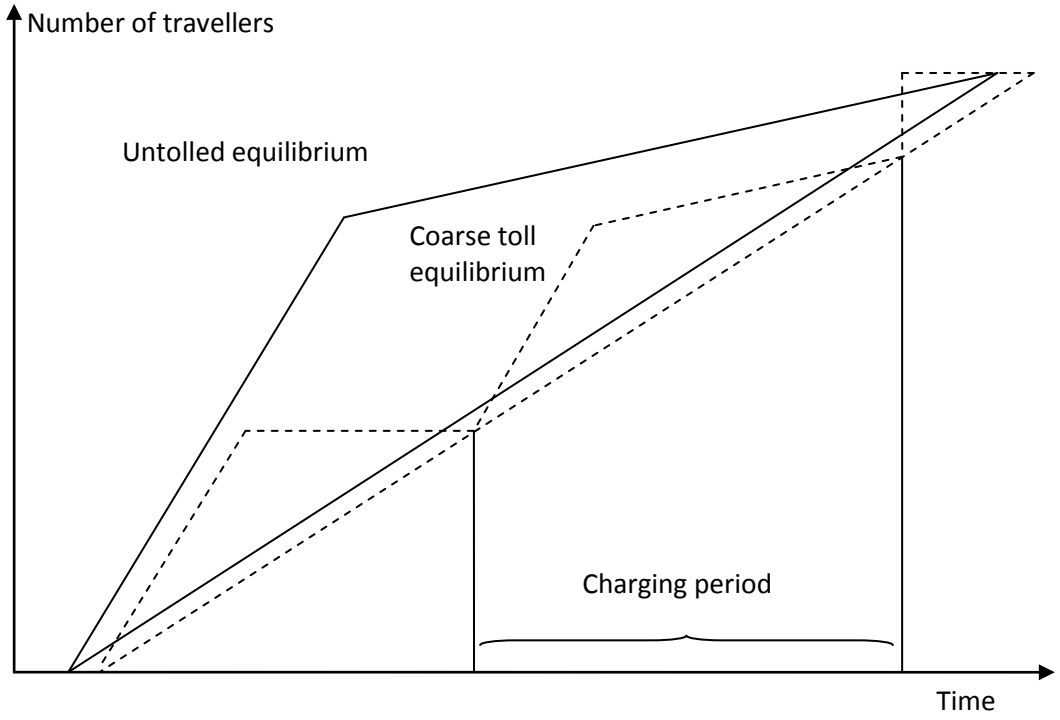


Figure 6 The optimal coarse toll

ADL (1990) carried out their analysis for the case of a single step coarse toll. Laih (1994; 2004) extended this analysis to the case of multistep tolls using a slightly modified queueing technology in which some travelers can wait in a separate queue for the toll to lift, while those paying the toll pass

the bottleneck.[9]Laih then showed that at most n/(n + 1) of the total queueing time can be eliminated with the optimal n-step toll. Daganzo& Garcia (2000) also consider a step toll with the modified queueing technology. They divide travellers into two groups. Travellers from the first group are not liable to pay any toll. Travellers from the second group are liable to pay a constant step toll if they pass the bottleneck during the tolling period, otherwise they do not have to pay any toll. If the toll is high enough then travellers from the tolled group will avoid the tolling period. The tolling period is timed such that it fits exactly with the equilibrium departure interval of the untolled group. As a consequence, travellers from the untolled group can find an equilibrium during the tolling period and be strictly better off than without the scheme. Travellers from the tolled group are not worse off, since they travel during the same interval as without the scheme and avoid paying any toll by travelling outside the toll period. The essential insight is that the equilibrium cost is determined by the first and last travelers (as in (5)or (11)) as long as capacity is fully utilized during the departure interval.

The function of the toll in this example is to reserve the bottleneck capacity for a specific group of travellers during a specific interval of time. Shen and Zhang (2010) describe a mechanism that uses ramp metering to achieve a similar effect.

5.2. Random capacity and demand

ADL (1999) consider bottleneck congestion in a situation where capacity varies randomly from day to day. The ratio is fixed within a day and given the ratio the evolution of the queue is then deterministic. Travellers choose departure time without knowing the random ratio of the day. They are assumed to find equilibrium in expected utility given the information they have. ADL identify circumstances in which the static model is not consistent with a reduced form of the dynamic model. A perhaps surprising result is that providing more information can decrease welfare when demand is elastic and congestion is not efficiently tolled.

Lindsey (2009) considers self-financing in the bottleneck model with random capacity and demand. He finds that the Mohring-Harwitz self-financing theorem survives randomness as long as the information used to set the optimal toll is the same as the information that is available to travellers.

de Palma and Fosgerau (2009) include random travel time variability in a different way. They consider the bottleneck model with fixed capacity but where the FIFO property of the bottleneck model is replaced by random queue sorting, where all travellers in the queue at any given time have the same probability of exiting the queue at that moment. A range of intermediate regimes is also considered. Equations (10) and (11) still apply and the results that follow from these hence also apply.

Queues take time to dissipate. This physical property of queues has implications for how queues evolve over the course of a day. An empirical regularity of congested demand peaks is that the mean travel time peaks later than the variance of travel time. Fosgerau(2010) shows how this phenomenon arises in a dynamic model of congestion with the ratio of demand to capacity being random.

5.3. Heterogeneity

An extension to the basic bottleneck model which is clearly very important is to allow for heterogeneity. The basic model describes travellers as having identical scheduling preferences and identical preferred arrival time. This is very far from reality. For example, using survey data, Fosgerau (2006) estimates the distribution of the value of travel time, α . After conditioning on a number of controls, he finds that the remaining variation in the value of travel time has more than a factor of 50 between the 20th and 80th percentiles of the value of travel time distribution. There is every reason to think that preferences regarding earliness and lateness are similarly heterogeneous.

[9]Laih (1994) did not recognize that it was necessary to reformulate the queueing technology in order to obtain his results. This was rectified in Laih (2004).

One of the first questions to ask when such heterogeneity is allowed in the bottleneck models is whether equilibrium still exists and whether it is unique. Analysis of the model would be severely complicated if this failed. This is the subject of Lindsey (2004), who presents general conditions under which equilibrium exists in the basic bottleneck model extended with heterogeneity in the form of a finite number of homogenous groups of travelers. Lindsey provides a review of previous literature regarding preference heterogeneity in the dynamic model.

5.4. Parking

Parking is costly in that it competes for urban space with other uses. Cruising for parking is a significant contributor to urban congestion. Arnott and co-authors have published a series of papers on this and related issues, a recent reference is Arnott and Rowse (2009).

There are a few papers on downtown parking in a dynamic framework in which parking occupies space and the attractiveness of a parking space decreases with the distance to the CBD. ADL (1991) use the bottleneck model to assess the relative efficiency of road tolls and parking fees. Without pricing, drivers occupy parking in order of increasing distance from the CBD. A time-varying toll can prevent queueing, but does not affect the order in which parking spots are taken. Optimal location-dependent parking fees may be superior; they do not eliminate queueing, but induce drivers to park in order of decreasing distance from the CBD, thereby concentrating arrival times closer to work start times. Zhang, Huang and Zhang (2008) integrate AM and PM commutes with parking in this framework.

5.5. Small networks with dynamic congestion

This section considers some simple extensions from one link to small networks. Consider first two routes in parallel connecting an origin with a destination. There are $N > 0$ travellers with $\alpha - \beta - \gamma$ scheduling preferences. They each have to choose a route and a departure time. Each route has a certain fixed travel time and a bottleneck with fixed capacity. Denote the fixed travel times by T_i and the capacities by $s_i, i = 1,2$. Denote also the number of travellers choosing route i as $n_i > 0$ where $N = n_1 + n_2$, since all travellers choose one and only one route. Moreover, let ρ_i denote the arrival rate at the bottleneck for each of the routes.

Consider first the choice of departure time conditional on the number of travellers on each route. From the previous analysis we know that in equilibrium they incur a trip cost of $\delta n_i / s_i$ on each route. There exists a unique equilibrium where

$$\alpha T_1 + \delta \frac{n_1}{s_1} = \alpha T_2 + \delta \frac{n_2}{s_2}.$$

This is equivalent to (1) for the static model. It is straightforward to verify that the equilibrium number of travellers on route 1 is

$$n_1 = N \frac{s_1}{s_1 + s_2} + \frac{\alpha}{\delta} \frac{s_1 s_2}{s_1 + s_2} \left(T_2 - T_1 \right),$$

and the equilibrium cost is

$$C = \alpha \frac{s_1 T_1 + s_2 T_2}{s_1 + s_2} + \delta \frac{N}{s_1 + s_2}.$$

This shows that two bottlenecks in parallel act just like a single bottleneck. The equivalent single bottleneck would have a fixed travel time that is a weighted average of the fixed travel times on the

two routes and it would have a bottleneck capacity that is the sum of the capacities of the two routes. This result can be generalised to any number of parallel routes.

A toll may be set at each bottleneck just as if it was a single bottleneck with elastic demand. As we have seen (page 22), the optimal toll does not affect the cost of using each route. Hence the split of travellers between routes is not affected by optimal tolling: The optimal toll does not reallocate between routes, but only across departure times. This is a very different conclusion than was reached in the static model, where the social optimum had a different allocation of travellers on routes than the equilibrium.

There is another situation in which several bottlenecks acts like a single bottleneck. This happens when bottlenecks are connected in a serial manner. In this case, the effective capacity is just the minimum of the bottleneck capacities. That is, the binding capacity constraint is that of the smallest bottleneck.

The property that parallel or serial bottlenecks can be reduced to a single equivalent bottleneck seems likely to survive if $\alpha - \beta - \gamma$ preferences are replaced by general preferences. The description of the equivalent bottleneck does become more complicated. The property that equilibrium usage of the parallel routes is optimal also survives.

ADL (1993b) analyze a Y-shaped network of bottlenecks to show that a Braess type paradox can arise: an increase in capacity can lead to increased cost. Analysis of more complicated networks is complicated and no general results on networks of bottlenecks seem to be available.

5.6. Large networks

The extension of the dynamic model to large networks remains a difficult problem. So far, existence and uniqueness of equilibrium have not been established (in spite of many attempts). The dynamic traffic assignment problem (Merchant &Nemhauser, 1978) is the subject of a large literature spanning several disciplines. Heydecker& Addison (2005) and Zhang & Zhang (2010) derive some analytical results.

Otherwise, the literature mostly uses numerical methods. Dynamic traffic assignment models are also difficult to work with numerically due to the dimensionality of the problem, which quickly becomes extreme.

Consider a simulation model in which travellers choose the least-cost path through a network. Conditional on the actions of all other travelers, the problem of finding the least-cost path is feasible to solve using well-established algorithms (Dijkstra, 1959). These algorithms are quite efficient but nevertheless require nontrivial time to execute. The dynamic version of such a model is formulated in continuous time; we may want to approximate it using discrete time steps of one second. In, say, a four hour peak period there are 14,400 possible departure time choices. In order to simulate the choice of departure time, we have to find the least-cost path for each possible departure time. Consider a city which can be adequately represented by a zone system of 500 zones. Then the OD-matrix, indicating the size of origin-destination flows is a 500 by 500 matrix with 25,000 entries. So the model will have to solve 3.6 billion shortest path problems through the network connecting the 500 zones. This will have to be done many times in order for such a simulation to identify an equilibrium in which no traveller will want to change his choice of departure time and route. The result is a huge computational problem and it is practically impossible to handle using a naive approach.

This section describes one approach taken to this problem, used in the model METROPOLIS (de Palma *et al.*, 1997). The basic idea for reducing the amount of computation is to drop the assumption that travellers can choose the shortest path considering the whole network at once. At each intersection, travellers are able to observe the travel cost on each downstream link. But they do not observe the travel cost on links further downstream. Instead they are able to form an expectation regarding the travel cost from the next downstream nodes until the destination. Travellers then

choose the next link with the smallest expected total cost to reach the destination, *i.e.* the smallest sum of the cost of the next link and the downstream expected cost. This portrays travellers as making dynamic discrete choices and these are readily formulated as a dynamic programming model using the Bellman principle.'[10]

The simulation model looks for equilibrium using a process which can be interpreted as a day-to-day learning process. At the end of each day, the past outcomes for all travellers are pooled and this pool of information is common knowledge. During the next day, travellers have this information available when forming expectations. The idiosyncratic error terms are the same day after day (for departure time choice model). The choice of route can be either deterministic or stochastic (in such case, error terms are i.i.d. over space and time).

5.7. Other congestion functions

Henderson (1974) formulated a dynamic model of congestion using a similar setup to Vickrey (1969), but in which the travel time is determined by the flow at the time of departure and where flows departing at different times do not interact. Chu (1995) showed that the original Henderson formulation had problems due to nonexistence of equilibrium and proposed a reformulation in which travel time for a traveller is instead determined by the flow at the time of arrival at the destination. The Chu formulation has the Vickrey bottleneck as a limiting case.

6. Conclusions

This paper has presented an overview of dynamic models of congestion, focusing on results derived from the Vickrey bottleneck model. This model combines in a compact way the essential features of congestion dynamics. We have also argued that some fundamental features of congestion are inherently dynamic, which makes dynamic models indispensable for many purposes. In particular, dynamic models can be used to study a variety of policies that cannot be studied with static models. These include road pricing with a time-varying component, flexible work hours, staggered work hours, dynamic access control, and ramp metering used to differentiate capacity allocation. Pricing policies are much more effective when tolls depend on the time of the day, for stylised as well as for real networks (see Santos, 2004).

Research into congestion dynamics remains a very active area with many unresolved issues of high importance. We will mention a few here. Economic analyses using dynamic models of congestion are usually undertaken on the assumption that users are in Nash equilibrium. It would therefore be of interest to give general conditions under which Nash equilibrium exists (for general networks). It would further be of interest to specify learning mechanisms that would lead to Nash equilibrium. A learning mechanism is a rule that travellers use to update their choice of departure time and route in the presence of information concerning past outcomes. The existence of learning mechanisms leading to Nash equilibrium would support the presumption that the notion of Nash equilibrium is useful as a benchmark for actual congestion phenomena. Knowledge about learning mechanisms leading to Nash equilibrium may also be useful for the design of algorithms to find Nash equilibrium in simulation models.

Progress would also be desirable concerning the nature of scheduling preferences. The discussion in this chapter has taken for granted that travellers are equipped with scheduling preferences and that these can be regarded as exogenous from the point of view of our analysis. Our transportation perspective has led us to be concerned with the timing of trips and we view travellers simply as having preferences regarding timing such that they can respond to circumstances by changing their trip timing in sensible ways. These times are hardly the fundamental objects of preference and, strictly speaking, it only makes sense to formulate preferences in these terms when circumstances such as the activities before and after the trip can be regarded as exogenous. This is, however, not a

[10] Dynamic discrete choice models are surveyed in Aguirregabiria& Mira (2010).

very appealing position. If I know that my trip will take more time, then I will adjust my schedule for the day to take this into account. I care, *e.g.*, about not being late for appointments. But I make the appointments myself and so my scheduling preferences are a consequence of choice (see the chapter of Pinjara and Bhat on this issue).

It is natural to ask why commuters mostly prefer to arrive at work at the same time. Various contributions have answered this question by pointing to agglomeration forces at the workplace, whereby productivity and wages are affected by the degree of overlap in work times, see Henderson (1981), Wilson (1988), Hall (1989). If this view is correct, then changes to the transport system will affect agglomeration, which in turn will affect commuter scheduling preferences. It remains to be seen how such mechanisms matter for our understanding of the effect of transport policies.

Endogeneity of scheduling preferences may also matter for the value of information. Consider a trip exposed to random travel time variability. At some point in time I will learn the size of delay. If scheduling preferences are exogenous, then it only matters whether I learn about the size of the delay soon enough to adjust my departure time. If scheduling preferences are endogenous, then it also matters whether I learn about the size of the delay soon enough to adjust my schedule (Kreps and Porteus (1978) consider dynamic choice behaviour under conditions of uncertainty, with emphasis on the timing of the resolution of uncertainty.

As discussed in this paper, the current state of the topic of dynamic congestion modelling provides a range of general insights from small stylised models. Numerical simulation models exist to deal with the complexities of real size networks. In between, there is a large gap. Numerical simulation has the drawback that it must rely on particular assumptions, which may or may not provide good approximations to the object of interest. So a main motivation for continued theoretical research into dynamic models of congestion is the desire for increased generality. The fewer assumptions required for a conclusion, the more certain we can be that it applies. As this chapter has discussed, there are a number of directions in which we would like to extend our models so that they become better able to account for the facts that travellers are very heterogeneous, they make route and scheduling decisions based on limited information, they interact heavily in ways related to scheduling and they move about in complex networks that are subject to random shocks. The other main motivation for research into the area is the potential for providing a better empirical foundation for our models. One possibility that naturally comes to mind is to seek to utilise data sources such as GPS data to obtain a better understanding of actual trip scheduling behaviour.

In conclusion, many exciting things have been done, giving us many important insights into congestion dynamics, and there are still many exciting things waiting to be done.

References

Aguirregabiria, V. & Mira, P. (2010), Dynamic discrete choice structural models: A survey,*Journal of Econometrics, 156,* 38-67.

Arnott, R. A., de Palma, A., & Lindsey, R. (1990), Economics of a bottleneck, *Journal of Urban Economics, 27,* 111-130.

Arnott, R.A., de Palma, A., & Lindsey, R. (1991), A temporal and spatial equilibrium analysis of commuter parking, *Journal of Public Economics, 45,* 301-335.

Arnott, R. A., de Palma, A., & Lindsey, R. (1993a), 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. (1993b), Properties of Dynamic Traffic Equilibrium Involving Bottlenecks, Including a Paradox and Metering, *Transportation Science, 27,* 148-160.

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.

Arnott, R. A. &Rowse, J. (2009), Downtown parking in auto city, *Regional Science and Urban Economics, 39,* 1-14.

Beckmann, M. J., McGuire, C. B., & Winston, C. B. (1956), *Studies in the Economics of Transportation*, New Haven, Connecticut: Yale University Press.

Chu, X. (1995), Alternative congestion pricing schedules, *Regional Science and Urban Economics, 29,* 697-722.

Daganzo, C. F. & Garcia, R. C. (2000), A Pareto Improving Strategy for the Time-Dependent Morning Commute Problem, *Transportation Science, 34,* 303-311.

de Palma, A., Ben-Akiva, M., Lefevre, C., & Litinas, N. (1983), Stochastic Equilibrium Model of Peak Period Traffic Congestion, *Transportation Science, 17,* 430-453.

de Palma, A. & Fosgerau, M. (2009),Random queues and risk averse users, *Working Paper*, EcolePolytechnique, France.

de Palma, A., Lindsey, R., & Proost, S. (2007), Investment and the use of tax and toll revenues in the transport sector: The research agenda, In *Investment and the Use of Tax and Toll Revenues in the Transport Sector* (pp. 1-26).

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

Dijkstra, E. W. (1959), A note on two problems in connection with graphs, *NumerischeMathematik, 1,* 269-271.

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

Fosgerau, M. (2009), The marginal social cost of headway for a scheduled service, *Transportation Research Part B: Methodological, 43,* 813-820.

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. & Engelson, L. (2010), The value of travel time variance, *Transportation Research Part B, Forthcoming*.

Fosgerau, M. &Karlstrom, A. (2010), The value of reliability, *Transportation Research Part B, 44,* 38-49.

Hall, R. E. (1989), Temporal agglomeration, *National Bureau of Economic Research, 3143*.

Haurie, A. &Marcotte, P. (1985), On the relationship between Nash-Cournot and Wardropequilibria, *Networks, 15,* 295-308.

Henderson, J. V. (1974), Road congestion : A reconsideration of pricing theory, *Journal of Urban Economics, 1,* 346-365.

Henderson, J. V. (1981), The economics of staggered work hours, *Journal of Urban Economics, 9,* 349-364.

Heydecker, B. G. & Addison, J. D. (2005), Analysis of Dynamic Traffic Equilibrium with Departure Time Choice, *Transportation Science, 39,* 39-57.

Kreps, D. M. &Porteus, E. L. (1978), Temporal Resolution of Uncertainty and Dynamic Choice Theory, *Econometrica, 46,* 185-200.

Laih, C. H. (2004), Effects of the optimal step toll scheme on equilibrium commuter behaviour, *Applied Economics, 36,* 59-81.

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

Lindsey, R. (2004), Existence, Uniqueness, and Trip Cost Function Properties of User Equilibrium in the Bottleneck Model with Multiple User Classes, *Transportation Science, 38,* 293-314.

Lindsey, R. (2009), Cost recovery from congestion tolls with random capacity and demand, *Journal of Urban Economics, 66,* 16-24.

Merchant, D. K. &Nemhauser, G. L. (1978), A Model and an Algorithm for the Dynamic Traffic Assignment Problems, *Transportation Science, 12,* 183-199.

Mohring, H. &Harwitz, M. (1962), *Highway Benefits: An Analytical Framework*, Evanston, Illinois: Northwestern University Press.

Santos, G. (2004), Research in Transportation Economics, 9, XI-XIII.

Shen, W. & Zhang, H. M. (2010), Pareto-improving ramp metering strategies for reducing congestion in the morning commute, *Transportation Research Part A: Policy and Practice, 44,* 676-696.

Small, K. (1982). The scheduling of Consumer Activities: Work Trips,*American Economic Review, 72,* 467-479.

Tseng, Y. Y. &Verhoef, E. T. (2008), Value of time by time of day: A stated-preference study, *Transportation Research Part B: Methodological, 42,* 607-618.

Verhoef, E. T. &Mohring, H. (2009), Self-Financing Roads, *International Journal of Sustainable Transportation, 3,* 293-311.

Vickrey, W. S. (1969), Congestion theory and transport investment, *American Economic Review, 59,* 251-261.

Wardrop, J. G. (1952), Some Theoretical Aspects of Road Traffic Research, *Proceedings of the Institute of Civil Engineering, Part II,* 325-378.

Wilson, P. W. (1988), Wage variation resulting from staggered work hours, *Journal of Urban Economics, 24,* 9-26.

Zhang, X. & Zhang, H. (2010), Simultaneous Departure Time/Route Choices in Queuing Networks and a Novel Paradox, *Networks and Spatial Economics, 10,* 93-112.

Annex B: Traffic Congestion Pricing Methodologies & Technologies

Refrence:

de Palma, A. and R. Lindsey (2011), Traffic Congestion Pricing Methodologies and Technologies. *Transportation Research Part C: Emerging Technologies*, 19(6), 1377-1399.

Abstract

This paper reviews the methods and technologies for congestion pricing of roads. Congestion tolls can be implemented at scales ranging from individual lanes on single links to national road networks. Tolls can be differentiated by time of day, road type and vehicle characteristics, and even set in real time according to current traffic conditions. Conventional toll booths have largely given way to electronic toll collection technologies. The main technology categories are roadside-only systems employing digital photography, tag and beacon systems that use short-range microwave technology, and in vehicle-only systems based on either satellite or cellular network communications. The best technology choice depends on the application. The rate at which congestion pricing is implemented, and its ultimate scope, will depend on what technology is used and on what other functions and services it can perform.

1. Introduction

Traffic congestion is common in large cities and on major highways and it imposes a significant burden in lost time, uncertainty, and aggravation for passenger and freight transportation. The European UNITE project estimated the costs of traffic congestion in the UK to be 15 billion/yr. ($23.7 billion/yr.) or 1.5% of GDP (Nash *et al*., 2003).[11] For France and Germany the estimates were 1.3% and 0.9% of GDP respectively. The Texas Transportation Institute conducts an annual survey of traffic congestion in US urban areas.[12] According to the 2009 report, in 2007 congestion caused an estimated 4.2 billion hours of travel delay and 2.8 billion gallons of extra fuel consumption with a total cost of $87 billion (Schrank and Lomax, 2009). With a US GDP of $14.1 trillion in 2007 the cost amounted to 0.6% of GDP.[13] The average cost per traveler in the urban areas studied was $757.

Most of the costs of traffic congestion are borne by travelers collectively, but because individual travelers impose delays on others they do not pay the full marginal social cost of their trips and therefore create a negative externality. The standard economic prescription to internalize the costs of a negative externality is a Pigouvian tax. In the first edition of his textbook, *The Economics of Welfare*,Pigou (1920) himself argued for a tax on congestion and thereby launched the literature on congestion pricing. Most economists have supported congestion pricing although many have been concerned about the details of implementation (Lindsey, 2006). Congestion pricing has a big advantage over other transportation demand management policies in that it encourages travelers to adjust all aspects of their behavior: number of trips, destination, mode of transport, time of day, route, and so on, as well as their long-run decisions on where to live, work and set up business.

For decades congestion pricing remained largely an ivory-tower idea, but interest gradually spread outside academia and congestion pricing has come into limited practice. The main operating schemes are High Occupancy Toll (HOT) lane facilities in the US, the London congestion charge, the Stockholm cordon charge[14], and Singapore's Electronic Road Pricing system. Few cost-benefit analyses of these (or other) congestion pricing systems have been undertaken. However, the limited evidence suggests that well-designed schemes can yield significant net economic benefits. Small et al. (2006) estimate the benefits from tolling a two-lane facility similar to the State Route 91 (SR-91) HOT lanes in Orange County, California.[15] Optimal tolling of both lanes yields a welfare gain of nearly $3 per trip, while operating one lane as a HOT lane and leaving the other lane free yields a still appreciable gain of $2.25 per trip.

The London congestion charge has been closely monitored since it was introduced in 2003. The fifth annual report (Transport for London, 2007) estimated the gross annual benefits of the original scheme at £200 million ($316 million) and the total costs at £88 million ($139 million), resulting in a net benefit of £112 million ($177 million) and a benefit-cost ratio of 2.27.[16] Stockholm's congestion

[11] Throughout the paper foreign currency amounts are converted to US dollars at December 12, 2010 exchange rates.

[12]Since June 2008, congestion in US cities has also been tracked by INRIX using data from GPS-enabled probe vehicles (http://www.inrixtraffic.com).

[13] http://www.bea.gov/national/index.htm#gdp, page B19.

[14]According to Swedish law the congestion charge is a tax, but it will be called a charge or toll in this review.

[15] HOT lanes run in parallel with lanes that are not tolled. High Occupancy Vehicles (HOVs) can use the HOT lanes without paying. The occupancy requirement is usually either two people (HOV2+) or three people (HOV3+) as on SR-91.

[16] This information is taken from Santos (2008). The congestion charging zone was extended to the west in 2007, but the Western Extension will be terminated in January 2011. Santos and Fraser (2006) determined that the Western Extension fails a cost-benefit test.

charge began as a seven-month Trial in 2007 and, after a successful referendum, became permanent in 2007. Based on results of the Trial, Eliasson (2009) estimated the annual benefits net of operating costs to be about SEK 650 million/year ($94 million) and investment and startup costs of about 1.9 billion SEK ($275 million) yielding a social surplus pay-off time of only four years. Singapore's Electronic Road Pricing has not been put to a comprehensive cost-benefit test, but the system is widely held up as a successful model.[17]

Several countries have considered regional or national road-pricing schemes — in part to internalize congestion and other traffic externalities.[18] However, despite the apparent success of existing schemes, and plans to establish more, congestion pricing continues to be a hard sell. Several major proposals have recently been scuttled by public or political opposition. Cordon tolling schemes for Edinburgh and Manchester were rejected by public referenda in 2005 and 2008, respectively. An online petition to the UK government in early 2007 attracted more than 1.8 million signatures against road pricing, and effectively put an end to plans for a national scheme in the UK for the time being. A cordon toll plan for New York City was stopped by the New York state legislature in April 2008 when it declined to vote on the proposal. And a plan to introduce a national distance-based charge in the Netherlands has been put on hold after the Dutch government collapsed in February 2010.

These setbacks illustrate the difficulties of designing congestion pricing schemes that are both efficient and publicly acceptable. Much has been written recently about road pricing in general, and congestion pricing in particular, and it is useful to delineate the bounds of this review as well as to provide a few references for material that is not covered. As the title of the review indicates, it concerns ways in which congestion pricing can be implemented and the technologies available for doing so. Considerable attention is given to comprehensive distance-based pricing because it appears to offer substantial potential benefits while also posing technological challenges.

Due to space limitations a number of topics related to congestion pricing are excluded from the review including parking congestion and parking pricing[19], pricing of road emissions[20], the use of congestion pricing revenues[21], and the role of congestion tolls in guiding efficient investments.[22]Slot-based reservation systems in which drivers book trips in advance are ignored[23] as are pricing instruments that may reduce congestion but are not designed to do so such as fuel taxes, vehicle

[17] The Milan EcoPass, which began operating in 2008, is designed primarily to reduce pollution.Rotaris et al. (2010) determine on the basis of the first eleven months of operation that EcoPass yields a net annual benefit.

[18] The European heavy goods vehicle (HGV) charging schemes described later in this review are designed primarily for revenue generation and for efficient and fair charging of long-distance road freight transportation. A study commissioned by Transport and the Environment (2010) finds that HGV kilometers are quite sensitive to charges with a central price elasticity of about -0.9. Much of the response is attributable to more efficient operation and chains of distribution rather than modal shift to rail or reduction in product demand. The study concludes that charges are likely to be effective in reducing road freight transport externalities.

[19]See Shoup (2005), Arnott et al. (2005, Chapter 2) and Arnott (2011).

[20] See Johansson-Stenman and Sterner (1998) and Jensen-Butler et al. (2008).

[21]See De Palma et al. (2007).

[22]See Small and Verhoef (2007, Chapter 5).

[23] See Wong (1997).

ownership taxes, vehicle registration fees, and Pay-As-You-Drive (PAYD) insurance[24]. Public-choice and other institutional considerations are mentioned only incidentally.[25]

The balance of the review is organized as follows. Section 2 provides a brief summary of the theory of congestion pricing with an emphasis on practical complications. Section 3 describes methods of congestion pricing as defined by network coverage and how tolls are differentiated by time of day, type of road, and other dimensions. Section 4 describes technologies that are used, or being tested, for congestion pricing, and reviews their strengths and weaknesses. Concluding remarks are made in Section 5.

2. Theory of congestion pricing

Although this review is primarily concerned with the methods and technologies for congestion pricing it is useful to begin by summarizing the theory in order to identify the functional requirements of an effective congestion pricing scheme.[26] Following Walters (1961) consider first the static model of tolling a single road link. Let Q denote flow on the link measured in vehicles per hour, and $c(Q)$ the generalized cost of a trip on the link (i.e. vehicle operating cost plus travel time cost). The total cost of Q trips per hour is then $TC = c(Q)Q$, the marginal social cost of a trip is $MSC = dTC/dQ = c(Q)+c'(Q)\,Q$, and the marginal external cost is $MEC = MSC - c(Q) = c'(Q)\,Q$. The Pigouvian toll is therefore $\tau = c'(Q)\,Q$. Since the equilibrium value of Q depends on the toll, it cannot be deduced without knowing the demand curve for trips. Nevertheless, if the toll is set iteratively as a function of the observed flow using a suitable day-to-day adjustment procedure, the system will converge to the system optimum.[27]

The Pigouvian toll formula for a single link extends to each link of a road network if all links can be tolled efficiently. Let a denote a link (or arc) in the network, Q_a the flow on link a, and c_a the generalized travel cost on link a. As Yang and Huang (1998) show, if c_a is independent of flows on other links the Pigouvian toll on link a is simply $\tau_a = c_a'(Q_a)Q_a$, $a \in A$, where A is the set of all links. The toll is a function of flow on the link, but it is independent of travel conditions on other links so that only local information is required to set the toll.[28] Moreover, because tolls are link-based rather than path-based, information is not required about the paths that vehicles follow through the network. This is advantageous in terms of both practicality and privacy since there is no need to track trip origins, destinations, or routes.

---

[24]See Proostand Van Dender (1998),Parry (2005), andBordoff and Noel (2008).

[25]Governance is discussed by Sorensen and Taylor (2005), the potential role of the private sector in building and operating toll roads by Gómez-Ibáñezand Meyer (1993) and Small (2008), public acceptability by Schade and Schlag (2003) and Trannoy (2011), and equity by Ecola and Light (2009).

[26]More comprehensive reviews of the theory are found in Lindsey and Verhoef (2001), Small and Verhoef (2007, Chapter 4),Tsekeris and Voß (2008)and Parry (2009).

[27] Yang et al. (2004) show how tolls can be set by trial and error to reach a system optimum without knowledge of demand functions. Zhao and Kockelman (2006) extend the idea to users with heterogeneous values of travel time who choose between routesaccording to random-utility maximization. Yang and Szeto (2006) propose an algorithm that converges, andYang et al. (2007) develop a steepest descent method that converges monotonically and quickly to the system optimum on a general network.

[28]In practice, the cost on one link often depends on flows on other links due to conflicting traffic movements at intersections, signal control, queue spillover, delays in passing on undivided highways, and so on. The Pigouvian tax is then equal to the inner product of the gradient of the link cost function and the vector of link flows (Hearn and Yildirim, 2002).

The simple Pigouvian theory bypasses many complications that have led to a rich and still expanding literature, but also make practical applications much more challenging than the simple theory suggests. Only some of the more important complications will be identified here.[29] One complication is that the value of travel time (VOT) that underlies the trip cost function $c(\ )$ is not a single number but rather an average that depends on the composition of users, which in turn varies with the level of the toll, by time of day, and other factors. Values of time can also depend on trip duration, and there is evidence that VOTs are higher under congested than uncongested travel conditions.[30]

A second complication is that traffic flows vary greatly by time of day, day of week, and season. Formulating a dynamic system optimum on a road network, deriving tolls that support the optimum, and solving the system of equations numerically remains a challenge despite many years of research.[31] Dynamic models include a supply-side that describes how traffic flows evolve over time and space, and a demand-side that describes how travel demand depends on time.[32]The oldest supply-side model is the hydrodynamic model in which speed is specified as a decreasing function of density, and speed, density and flow evolve according to a partial differential equation. The hydrodynamic model can be implemented numerically using the cell transmission model (Daganzo, 1994). An alternative approach, which is less general but analytically more tractable, is to assume that congestion takes the form of queuing at bottlenecks, and that in the absence of a queue vehicles can move at free-flow speeds.

On the demand-side most models follow Vickrey (1969) in assuming that travellers (individuals or freight transporters) have preferences for when they start or complete a trip, and incur a schedule delay cost if they travel at a non-ideal time. Each traveller chooses a departure time to minimize their generalized trip cost which includes vehicle operating costs, travel time, schedule delay cost, and tolls (if any). Deriving the system optimum on a network requires solving a difficult optimal control problem in which the departure rate and order of travelers for each origin-destination pair has to be solved as well as the routes they take (the dynamic traffic assignment problem). Arnott and Kraus (1998) show that in the case of a single link with first-in-first-out queue discipline the system optimum can be supported using a toll provided the toll can be varied freely over time. The toll incorporates both a static component analogous to the static Pigouvian toll, and a dynamic component. The toll does not necessarily rise and fall in sync with congestion, and its time evolution can be quite complex.[33]

---

[29]See Small and Verhoef (2007, Section 4.2) for further detail.

[30] Using stated preference data, Calfee and Winston (1998) find that disutility from congested travel time for commuting trips is three times as large as disutility from uncongested time. They conjecture that this is because commuters can relax while driving under uncongested, but not congested, conditions. From a meta-analysis of 143 British studies Wardman (2001) concludes that time spent in congested traffic is valued 48% more highly on average than time spent in free-flow traffic. He attributes this (p.112) to "more difficult driving conditions linked with greater stress, frustration and perhaps arrival time uncertainty". Small and Verhoef (2007, p.54) note that if travel time reliability is omitted, VOT estimates may reflect aversion to unreliability. This will lead to greater upward bias in VOT estimates under congested conditions if congestion and unreliability are positively correlated.

[31]See Carey and Srinivasan (1993), Ghali and Smith (1995), Boyce (2007), and Friesz et al. (2008).

[32]The summary here is based on Lindsey and Verhoef (2001). For a more recent survey of dynamic models see de Palma and Fosgerau (2011).

[33] If congestion takes the form of queuing, the toll can be solved analytically in certain cases.In all practical applications to date, tolls do not vary continuously but rather are adjusted intermittently in steps. Step tolls are not fully efficient, and they can induce drivers to speed up just before a toll is

A third complication in computing congestion tolls is that congestion varies not only predictably with recurrent demand patterns but also unpredictably due to accidents, bad weather, special events, transit strikes, and other shocks.[34] Tolls should therefore vary according to real-time conditions and they should reflect the value that travelers place on travel time reliability as well as on travelers' values of (average) travel time.[35]

A fourthcomplication in computing tolls is that the congestion externality a vehicle imposes depends on its size, acceleration and braking capabilities, and maneuverability. These factors are typically accounted for by using a Passenger Car Equivalent (PCE) factor (Transportation Research Board, 2000). The PCE of large vehicles is often adjusted to account for hilly terrain, but it can also depend on the proportion of large vehicles in the traffic stream, and a large vehicle can have asymmetric effects on light and heavy vehicles (Peeta et al., 2004).

A fifth complication is that first-best (i.e. Pigouvian) tolling is efficient only if all links can be tolled. If they cannot, tolls should be set following second-best principles. A widely-studied case is one with two parallel (i.e. substitute) links of which only one can be tolled. The toll is set below the marginal external congestion cost on the tolled link in order to limit congestion on the untolled link.[36] Analogously, if a link has an untolled complement such as a link upstream or downstream, the second-best congestion toll exceeds the marginal external congestion cost on the tolled link.

A sixth complication is that congestion affects the magnitudes of other road-traffic externalities including accidents (Hensher, 2006; Steimetz, 2008), emissions (Daniel and Bekka, 2000; Glaister and Graham, 2005) and road damage (Hussain and Parker, 2006). This interdependence would not matter for setting congestion tolls if the external costs of accidents, emissions, and road damage were internalized by efficient pricing or some other means, but since these costs are not fully internalized these knock-on effects should, in principle, be factored in when setting congestion tolls.

A seventh consideration is that externalities and other market failures arise not only in road transportation but also with other transport modes and in other economic sectors. For example, urban public transit service has scale economies (a positive externality), but it is also heavily subsidized in most cities, and depending on which influence is larger fares can be overpriced or underpriced. Labor markets are distorted by income taxes and this has implications for tolling commuting and work-related trips (Parry and Bento, 2001; van Dender, 2003). And traffic congestion affects agglomeration economies in urban areas (Graham, 2007). Levying congestion tolls could exacerbate, or ameliorate, these distortions and studies have shown that the effects may be of first-order importance.[37]

---

raised, or slow down before it is lowered. Nevertheless, step tolls become more efficient as the number of steps is increased (Laih, 1994, 2004; Lindsey et al., 2010).

[34] Nonrecurring traffic congestion accounts for a large fraction of total delays in major urban areas. According to Schrank and Lomax (2009, Exhibit A-9) incident-related delays on US freeways range from 70% to 250% of recurring delay in the 43 largest urban areas.

[35]Small and Verhoef(2007, Section 2.6) review the theory and estimation of the value of travel time reliability.

[36]See Small and Verhoef (2007, pp. 139-142).

[37]For example, Parry and Bento (2001) show that by discouraging labor force participation, a commuter tax can impose a welfare loss that exceeds the Pigouvian welfare gain from alleviating congestion. Graham and Van Dender (2008) analyze a model with workers who are employed in sectors with different degrees of agglomeration economies. They show that the potential welfare gains from tolling are severely diminished if tolls cannot be differentiated between sectors, or if toll revenues cannot be redistributed to workers according to sector.

This brief review of the theory of congestion pricing reveals that congestion tolls should be differentiated by vehicle type, road link, time of day, real-time traffic conditions, trip purpose, and local conditions such as pricing of public transit service and other substitute modes of transport. In practice, tolls cannot be freely varied along all these dimensions. For technological, economic, or public acceptability reasons it may not be possible to toll all roads, to adjust tolls frequently by time of day, or to vary tolls according to traffic conditions. Lack of information or legal prohibitions may also preclude toll discrimination according to certain vehicle or driver characteristics.

In principle, the complications listed above (and many others) should be weighed when choosing a congestion pricing scheme and the levels and structure of tolls. In practice this is infeasible at anything like the theoretical ideal. Nevertheless, if the various complications are simply ignored a congestion pricing scheme may perform badly, and quite possibly could be worse than doing nothing.[38] Care should therefore be taken in deciding which complications are too important to ignore in a given application.

Congestion pricing schemes can be categorized along several dimensions: (1) the type of scheme (e.g. facility-based, area-based, or distance-based), (2) the degree to which tolls vary over time, (3) other dimensions of toll differentiation, and (4) technology. Section 3 addresses the first three dimensions and Section 4 follows by discussing technology in more detail. This sequence is followed for two reasons. First, it facilitates presentation. Second, technology choice is subordinate to choices along the other three dimensions in the sense that technology is not an end in itself, but rather should be driven by policy needs. This does not imply, of course, that technology is unimportant. As discussed in Section 4, technology choice affects system infrastructure and operating costs, flexibility, scalability, ability to differentiate tolls and other features of road pricing schemes. Furthermore, no technology yet exists to implement the most sophisticated forms of congestion pricing that approach the theoretical ideal. Technology choices therefore cannot be left until decisions have been made on how to implement congestion pricing. In practice, the choices are likely to be made iteratively, and with repeated visits back to the “drawing board”.

3. Methods of congestion pricing

3.1. Types of congestion pricing schemes

Congestion pricing schemes can be classified in various ways. The four categories considered here are presented roughly in order of increasing scale. Their advantages and disadvantages are addressed later.

Facility-based schemes

For centuries tolls have been imposed on roads, bridges, and tunnels, and this is still the most common form of road pricing by far although tolls designed to price congestion have only been implemented on a few facilities. Tolls can be levied either on all lanes of a facility or on designated toll lanes as is done on HOT lane facilities. Tolls can also be levied either at a single point on a facility or at multiple points with the total amount paid determined by distance traveled (e.g. as on Highway 407 in Toronto and on the I-15 Express Lanes which opened in 2009).[39] In the US a relatively new concept called "Managed lanes" has been developed which combines tolls, vehicle eligibility restrictions, and access control to regulate demand (FHWA, 2008). (HOT lanes combine tolls and high occupancy vehicle restrictions.) The goal is to maintain optimal driving conditions (possibly free-flow speeds) in response to changing conditions.

[38] Examples include untolled substitutes, labor market distortions, and agglomeration economies as discussed above.

[39] See http://www.407etr.com/About/tolls.htm and http://fastrak.511sd.com/ [December 9, 2010].

Cordons

Toll cordons are a form of area-based charging in which vehicles pay a toll to cross a cordon in the inbound direction, in the outbound direction, or possibly in both directions. A cordon scheme can encompass multiple cordons, and it can include radial screenlines to control orbital movements. All existing schemes are single cordons. The Norwegian toll rings were the first cordons to be created, but their main purpose has been revenue generation rather than congestion pricing.[40] The Milan EcoPass, introduced in 2008, is designed primarily to reduce emissions; congestion relief is only a secondary objective (Rotaris et al., 2010).

The only cordon scheme designed to manage congestion is the Stockholm congestion charge.[41] The cordon surrounds the city center and has 18 control points. Tolls are paid on each inbound passage up to a daily maximum of 60 kronors ($8.68). Pricing is in effect on weekdays from 6:30 to 18:30. The toll is 10, 15, or 20 Swedish kronors ($1.45, $2.17, or $2.89) depending on time of day. There is no charge on weekends, holidays, or the day before holidays.[42]

Singapore's Electronic Road Pricing (ERP) scheme, launched in 1998, covers certain expressways and arterial roads as well as three restricted zones in the CBD and the Orchard cordon. It is therefore a hybrid of facility-based tolls and cordons. Tolls are generally varied every half hour. As in Stockholm, payment is required for each passage or entry.[43]

As noted in the introduction, cordon schemes for Edinburgh, Manchester and New York City have been proposed, but rejected. Nevertheless, others are being planned. Gothenburg, Sweden intends to introduce a cordon in 2013 with a toll structure similar to the one in Stockholm (Pustral UGM, 2010). And a cordon in the northeast portion of San Francisco with a weekday AM and PM peak-period toll is being studied (San Francisco County Transportation Authority, 2010).

Zonal schemes

With a zonal scheme (sometimes called an area charge) vehicles pay a fee to enter or exit a zone, or to travel within the zone without crossing its boundary. Zone boundaries can be defined by natural features such as rivers, lakes, oceans, and mountains, as well as by elements of the built environment such as roads, tunnels, bridges, residential neighborhoods, and jurisdictions (states or provinces can define zones). The only operational zonal congestion pricing scheme is the London congestion charge, introduced in 2003.The original charging zone comprised a 21 km^2 area around the city center. A flat charge of £5 per day was levied on weekdays from 7:00 -18:30 for driving anywhere within the zone or for parking on public roads. In 2005, the toll was raised to £8, and in 2007 the charging period was shortened to end at 18:00 and the charging zone was expanded to include residential neighborhoods to the west. The Western Extension will be abolished in January 2011, and the daily charge for the original zone will increase to £10.

---

[40] Toll rings were established in Bergen (1986), Oslo (1990), and Trondheim (1991) as well as Kristiansand, Stavanger, Namsos, and Tønsberg. The Trondheim cordon was converted to a multi-sector zonal scheme in 1996, but tolling ended in 2005 when the policy package that included the toll ring expired.

[41] Since 2002, a £2.00 ($3.26) charge has been levied on vehicles entering the centreof Durham, England. The scheme operates like a cordon although only one, narrow public access road is involved. See Santos (2004) and http://www.durham.gov.uk/Pages/Service.aspx?ServiceId=6370 [December 9, 2010].

[42] For details see Eliasson et al. (2009a).

[43] http://www.lta.gov.sg/motoring_matters/motoring_erp.htm [December 9, 2010]. From 1975 to 1998,Singapore operated an Area License Scheme to control traffic in the CBD. Despite its name, the license was only required for vehicles traveling into the charging zone, not within it, and it was therefore effectively a cordon rather than a zonal scheme.

Travel along the boundary of the charging zone is free. Several vehicle categories are exempt, and residents of the charge area receive a 90% discount.[44]

Distance-based schemes

With distance-based schemes, charges vary with distance travelled, either linearly or nonlinearly. As noted above, some facilities charge on the basis of distance. Networks of truck-only toll lanes and networks of HOT lanes are under consideration, and tolls on these networks are likely to be distance-based as well.[45] For schemes that encompass multiple roads or regions the charge rate can depend on type of road. Four US states have implemented distance or weight-based charges for heavy goods vehicles (Conway and Walton, 2009) but the charges are intended to recover the infrastructure costs imposed by heavy vehicles rather than to manage demand. National distance-based heavy goods vehicle tolls exist in Switzerland, Austria, Germany, the Czech Republic and the Slovak Republic, and other European countries are developing or considering them. For the purpose of this review these schemes are mainly of interest regarding the technologies they use which will be discussed in Section 4.

Degree of time differentiation

Pricing schemes in general — and road tolls in particular — can be characterized as flat, time-of-day, or responsive.[46]*Flat tolls* are constant over time. Historically, tolls on most facilities were flat because of technological or administrative difficulties in changing the toll. For some schemes the toll prevails 24 hours a day. For others, such as the London congestion charge, the toll is levied at a constant rate during daytime on weekdays and not levied at other times.

Time-of-day tolls vary by time of day, day of week, and season according to a predetermined schedule. Examples include some HOT lane facilities in the US, Singapore's Electronic Road Pricing, and Stockholm's congestion charge. The time intervals between toll adjustments vary across schemes, and in some cases the interval in a given scheme varies by time of day.

Responsive tolls vary in real time (or near real time) as a function of prevailing traffic conditions. The only examples of responsive pricing are a few HOT lane facilities where tolls are adjusted to maintain free-flow speeds. During the early 1990s a congestion pricing trial was conducted in Cambridge, UK, in which drivers paid a charge when travel speed dropped below a threshold value. The logic underlying this scheme was similar to responsive pricing on HOT lanes except that the Cambridge scheme applied to all roads within the central city zone.[47]

---

[44] http://www.tfl.gov.uk/roadusers/congestioncharging/ [December 8, 2010].

[45] Truck toll road networks are proposed in Samuel et al. (2002) and Poole and Orski (2003). The National Capital Region Transportation Planning Board conducted a study of tolling a highway network in the Washington, D.C.region (Eichler et al., 2008; Bansal and Smith, 2010). Several combinations of lane construction, conversion of High Occupancy Vehicle (HOV) lanes to HOT lanes, and introducing tolls on existing general-purpose lanes were considered. Threetolled segments of the network are being constructed.

[46] Terminology varies. Time-of-day pricing is often referred to as "variable" or "scheduled", and responsive pricing as "dynamic". Cottingham et al. (2007) use "staticpricing" to describe any fixed schedule of charges that is announced well in advance (i.e. before travel decisions are made), and "dynamic charging" to describe charges that depend on contemporaneous congestion. Staticpricing in their nomenclature covers both flat and time-of-day pricing as defined here, and dynamic charging corresponds to responsive pricing here.

[47] The Cambridge scheme failed to advance beyond the trial because of opposition to the form of pricing and the perception that congestion was not bad enough to warrant tolls.Oldridge (1995) describes the planning, politics, and technology behind the experiment.

Responsive tolls are "reactive" in the sense that they are set, with a short time lag, as a function of current congestion levels. Yin and Lou (2009) develop two approaches for setting tolls on toll lanes. One (a feedback-control approach) increases the toll if lane occupancy exceeds a target level. The other (a reactive self-learning approach) learns motorists' willingness to pay a toll and iteratively adjusts the toll to maintain free-flow conditions on the tolled lanes while maximizing throughput. Lou et al. (2011) build on Yin and Lou (2009) by using a more realistic representation of traffic dynamics and an explicit formulation for toll optimization.

A step beyond reactive pricing is an anticipatory, or predictive, scheme in which tolls are based on forecast congestion. Dong et al. (2007) develop an algorithm to implement predictive pricing on a HOT lane facility and show that it can anticipate breakdowns in flow and maintain higher throughput than reactive pricing. Predictive pricing has long been envisaged as a tool for traffic management, but the information, communications, and computational requirements are challenging.

Other dimensions of differentiation

As noted in Section 2, optimal congestion pricing calls for tolls to be differentiated in several dimensions besides time of day. Differentiation by vehicle type, number of axles, and weight is common practice although these vehicle characteristics are imperfectly correlated with the congestion externality that a vehicle imposes since the externality also depends on road characteristics and the mix of users on the road. Toll differentiation according to speed and other correlates of dangerous driving behavior has been precluded by lack of information until recently although technological advances in on-board computers now make such differentiation feasible, and possibly practical.

Toll discounts and exemptions for certain categories of vehicles and drivers are also common. Various categories are exempt from the London and Stockholm congestion charges. London offers a 90% discount to residents, a 12.5% discount to fleets, and various discounts for monthly and annual payments. A number of toll roads and HOT lane facilities also offer quantity discounts in the form of reduced prices for advance purchase of multiple trips, or discounts based on cumulative usage over an accounting period (Wang et al., 2011). Quantity discounts are commonly used in transportation markets[48] as well as other sectors of the economy, but they are inconsistent with congestion pricing according to marginal social cost pricing principles. In some cases the discounts are used as a way to boost revenues, and in other cases they are offered for public acceptability reasons.

3.2. Choice of congestion pricing scheme

Having described the main characteristics of congestion pricing schemes we now turn to the central problem of determining which type of scheme (if any) is best in a given setting. Most studies have focused on the design of a particular type of scheme rather than the choice between schemes although comparisons are becoming more common. Attention is focused in this section on the choice between facility-based schemes (with varying degrees of road network coverage) and area-based schemes. The most important consideration is whether a scheme can target congestion according to where, and when, it occurs without inducing excessive spatial or temporal traffic diversion.

As explained in Section 2, link-based congestion tolls are theoretically optimal when all links can be tolled and the tolls can be freely differentiated by link, time of day, vehicle type, and other relevant dimensions. Under these conditions, a facility-based congestion pricing scheme with individually-optimized tolls for every facility (link) of the road network would be optimal — at least before tallying toll collection costs. In practice, neither comprehensive tolling nor freely-differentiated tolls is likely to be feasible for some time. The degree of network coverage, and the scope for toll differentiation, are therefore key determinants of how well facility-based schemes perform.

[48] See Anderson and Renault (2011).

High Occupancy Toll lanes are the smallest-scale existing congestion pricing schemes. Tolls are only paid on part of the capacity of a single road, and high occupancy vehicles are exempt. Travelers can therefore avoid paying a toll by sharing a ride, by using the toll-free lanes, or by selecting another route if one is available. Simple models with two routes in parallel and identical users (e.g. Verhoef et al., 1996) indicate that the maximum potential benefits from congestion pricing are rather modest unless a large fraction of road capacity can be tolled. The economics improve somewhat when heterogeneity of driver preferences is taken into account (Verhoef and Small, 2004) and improve further when value of travel time reliability is factored in (Small et al., 2006). The fraction of the first-best efficiency gains that can be derived from partial network tolling is also higher with time-of-day tolls than with flat tolls (Braid, 1996; De Palma et al., 2004) because varying tolls over time to suppress congestion reduces the social cost of trips as well as the amount of traffic diversion onto untolled capacity.

Assessing the performance of partial tolling schemes on real road networks is complicated by the large number of origins and destinations; by differences in the capacities and lengths of links; and by network topology which creates a complex interdependence between flows — with some links effectively operating as substitutes and others as complements. May et al. (2008) compute optimal (static) congestion tolls on the Edinburgh road network for different numbers of tolled links (*n*) on the assumption that the set of links that are tolled can be chosen freely for each *n*. They find that the benefits from tolling increase at a declining rate with *n*, and conclude (p.149) that "less than 10 per cent of the links are required to achieve around 60-70 per cent of the first-best benefits". This conclusion contrasts with the lessons from the two-routes-in-parallel network mentioned above in which there are sharply increasing returns from tolling both routes rather than one. The discrepancy in results highlights the difficulty of drawing general conclusions about the benefits from congestion pricing on networks.

Besides failing to price congestion on all of the network, partial tolling has the drawback that it exacerbates congestion on untolled links and may cause safety and infrastructure damage too if untolled links are built to a lower design standard than the links that are tolled. The extent of traffic diversion depends on the availability of convenient, alternative toll-free links. Re-routing options are limited in some US cities such as Boston, San Francisco, and Seattle, as well as cities such as Stockholm. Traffic diversion has been a problem on some toll roads.[49] Experience varies with the Heavy Goods Vehicle (HGV) charging schemes in Europe. Traffic diversion has been more of a problem in Austria where tolls are imposed only on primary roads, than in Switzerland where tolls are levied on all roads (Sorensen and Taylor, 2005). The German HGV toll is limited to federal motorways and some secondary roads, but traffic diversion has been minimal because many potential alternate routes are either closed to trucks or significantly slower (Broaddus and Gertz, 2008).

Compared to facility-based pricing of individual roads or small-scale road networks, area-based schemes have an advantage in intercepting more trips and are generally less susceptible to traffic diversion.[50] As noted in the introduction, the London and Stockholm schemes appear to be economically beneficial although this has been disputed.[51] Before the Stockholm cordon charge began there was concern that traffic within the zone would increase because no charge is levied while traveling within it. Schemes with sub-zones (as in the former Trondheim toll ring) and

[49]Swan and Belzer (2008) describe a case of traffic diversion off the Ohio Turnpike.

[50]For example, Olszewski and Xie (2005) find that toll elasticities in Singapore are higher for expressways than for the city center cordon.

[51]Prud'homme and Bocarejo (2005) conclude that the London charge is not cost effective, andPrud'homme and Kopp (2006) make a similar claim about the Stockholm charge.Mackie (2005) critiques Prud'homme and Bocarejo's (2005)analysis.

differential charge rates were considered before settling on a single cordon with a single charge (Eliasson et al., 2009b).

One reason for the good performance of the Stockholm charge is that the city is built on islands and just 18 access points suffice to form a cordon around it. In urban areas without natural boundaries choosing the number of cordons and where to locate them can be difficult. May et al. (2008) assess a variety of cordon options for London that differ in the number of cordons, the direction of movement in which charges apply, and the inclusion of radial screen lines. They find that the best-performing option (with three cordons, four screen lines and bi-directional charges) yields several times the benefits from a single cordon. The best single cordon performs poorly because it imposes the same charge on all journeys and allows many journeys to escape payment by rerouting. This illustrates how differences between cities in topology (and perhaps other factors such as population, area, public transit service quality and so on) influence the effectiveness of particular types of schemes.

Zonal schemes share most of the advantages and disadvantages of cordons. Zones are superior insofar as drivers are charged for moving within the zone, although the charge is independent of distance and for very short trips a zero charge would be superior.[52] The choice between cordons and zonal schemes has not been extensively studied[53] and practical experience is limited since the London and Stockholm schemes are the only large-scale examples of each type that were designed for congestion pricing.

A frequently-asked question is whether area-based schemes of the sort implemented in Europe would work in the US. Several authors in Richardson and Bae (2008) address this question and their overall assessment is negative. Compared to Europe, congestion in the US is less concentrated in city centers and more prevalent on expressways. With the exception of New York City, no US metropolitan area experiences congestion as severe as London's. Urban sprawl and trip chaining undermine public transit as a viable alternative to driving in the US. Public demands that toll-free routes be available are also stronger in the US than in Europe, and this too militates against charging whole areas rather than selected facilities.

3.3. Choice of time variation

For each type of charging scheme reviewed in the previous section there is a choice between flat, time-of-day, and responsive tolls.

Flat tolls

It is often claimed, or implicitly assumed, that flat tolls are suitable for maximizing revenue whereas time-of-day or responsive tolls are preferable to control congestion. Reality is not quite as black-and-white. A revenue-maximizing toll road operator has an incentive to internalize the congestion costs that users impose on each other, and will therefore generally prefer a toll that varies over time. Furthermore, the price elasticity of demand generally varies over time (demand is often less elastic

[52] Mayor Bloomberg's initial plan for congestion pricing in New York Cityfeatured car tolls of $8 for entering or leaving the charging zone, and a lower $4 toll for driving inside it.

[53] De Palma et al. (2005a) use a dynamic simulator to compare cordon and zonal charges for an area bounded by a ring road on a stylized urban road network. The zonal charge is paid for more trips, and has a greater effect on mode choice and departure-time decisions. Since average trips within the area are shorter than trips to or from outside, the optimal zonal charge is lower than the cordon toll but it leaves a larger fraction of travelers worse off. Maruyama and Sumalee (2007) use a static network equilibrium simulation to compare cordon and zonal charges for chained trips in Utsunomiya city near Tokyo. They find that the zonal charge is slightly more inequitable than the cordon toll, but the difference in impacts is slight.

during peak than off-peak periods) and so, therefore, does the profit-maximizing markup that the operator will want to charge on top of the Pigouvian toll. In London, travel speeds in the city center were relatively constant during daytime hours prior to introduction of the charge, and little might have been gained by varying tolls during the charge period (Leape, 2006). All that said, given the ease of varying tolls using Electronic Toll Collection technology, either time-of-day or responsive tolling is generally preferable for congestion pricing under most circumstances.

Time-of-day tolls

Time-of-day toll schedules are defined by the level of the toll in each time step and the time intervals between steps. Most analytical studies assume that schedules are chosen to be second-best optimal (i.e. to maximize welfare subject to applicable constraints). Practice has been rather more pragmatic. In Singapore, toll schedules are adjusted quarterly, and during June and December school holidays, to maintain target speeds of 45-65 km/h on expressways and 20-30 km/h on arterials at least 85% of the time (Chew, 2008). On SR-91, tolls are adjusted every six months using information on traffic volumes to maintain free-flowing conditions on the Express Lanes without reducing their throughput.[54]

The toll-setting rules in Singapore and on SR-91 effectively apply first-best pricing rules to the tolled links while ignoring congestion on other links. Such a policy is generally inferior to second-best pricing and it can be worse than not tolling at all (Verhoef et al., 1996). However, De Palma et al. (2005b) show that with bottleneck queuing congestion a "no-queue" tolling policy to eliminate queues performs fairly well relative to second-best tolling. Moreover, setting tolls to maintain a target level of service has some advantages.[55] First, it is less computationally demanding than second-best tolling because only information about the tolled links is required. Second, as noted in Section 2, tolls can be found by trial and error and (as in Singapore and on SR-91) periodically adjusted as demand evolves or as capacity changes elsewhere on the road network. Third, the decision rule is readily explained to users and the general public, and fourth it is easy to verify using traffic flow data that the rule is being followed.

The time interval between toll steps varies across schemes. On SR-91 it is one hour, in Stockholm it is 30, 60 or 90 minutes during peak periods and longer during the middle of the day, and in Singapore tolls generally vary every half hour. Schedules with small time steps may be difficult to remember, but they have the advantage that tolls change by small amounts between steps and motorists have less incentive to speed up or slow down in order to catch the "lower" toll. In 2003, Singapore introduced five-minute graduated rates between some half hour periods for this reason (Chew, 2008).

Responsive tolls

Responsive tolls have only been implemented to date on a few HOT lanes. On I-15 in San Diego and I-394 in Minneapolis the goal is to maximize utilization of the toll lanes while maintaining free-flow speeds. Tolls are adjusted as frequently as every six minutes on I-15 and every three minutes on I-394. Responsive pricing has the advantage that tolls can be adjusted according to actual travel conditions. Thus, if an accident blocks a lane on a multi-lane highway, the toll can be raised in order to limit the number of drivers who enter the highway during the disruption.

Responsive tolling has worked well on HOT lanes, and it may be practical on individual facilities where all capacity is tolled.[56] But there are several caveats. First, responsive tolling would probably not be suitable for area-based schemes unless public transit and other alternatives to driving have

[54] http://www.octa.net/91_tollpolicies.aspx [December 9, 2010].

[55] For further discussion see De Palma et al. (2005b).

[56] Responsive pricing has been implemented on multiple consecutive roadway segments for the MnPASS I-394 HOT lane project (Halvorson and Buckeye, 2006).

adequate capacity to accommodate travelers who do not want to pay a high toll. Second, responsive tolling can be effective only if travelers are aware of tolls sufficiently far in advance for them to modify their travel decisions. Third, individuals may be risk-averse to uncertain charges (Bonsall and Knockaert, 2008; Lindsey, 2009). Indeed, some businesses were against responsive tolling on I-15 because it would create uncertainty about their monthly bills (Bonsall et al., 2007).

3.4. Scheme complexity

The efficiency of markets depends on how much consumers know about prices and how much effort they need to expend to obtain the information. This general economics principle applies to the use of roads and tolls. From a user's perspective the complexity of a congestion pricing scheme depends on how much tolls vary by type of road, location, and time of day; whether tolls are responsive; how total amount paid varies with distance driven, and so on. Are there discounts for purchasing multiple cordon passes? Are there ceilings on the amount paid per day? Does the charge paid depend on the method of payment?

The dangers of designing an overly complex price system are highlighted in the report of the US National Surface Transportation Infrastructure Financing Commission (2009, p.141):

> "Even a road pricing system … where the payment system does not change, entails new information about the costs of traveling at certain times and on certain roads. This requires people to know more and to make more informed and more frequent decisions about travel."

If travelers are misinformed about tolls, they are liable to make mistakes that leave them worse off and that also undermine overall system efficiency because their responses deviate from what is intended. As travelers become accustomed to a charging system they may err less frequently, but they may also fall into habits and fail to modify their decisions if circumstances change. And if the system is very complex it may be strongly opposed. As Bonsall et al. (2007, p.680) remark: "A prime requirement is that the logic of the charge structure, and the necessity of a degree of complexity, is capable of being communicated and is seen to reflect the objectives of the scheme."

4. Congestion pricing Technologies

A number of studies have reviewed the technologies currently or potentially used for distance-based road user charges and congestion pricing: Sorensen and Taylor (2005), Cottingham et al. (2007), Ukkusuri et al. (2007), Broaddus and Gertz (2008), Bomberg et al. (2009), Donath et al. (2009), NCHRP (2009a), Noordegraaf et al. (2009), and Iseki et al. (2010). Several of the studies evaluate the relative merits of the technology options and their views will be mentioned in the conclusions.

4.1. Functions to perform and types of systems

All congestion pricing systems must perform three basic functions: (1) measurement of road usage by identifying vehicles and recording their locations and/or the distance they have traveled as well as their characteristics such as number of axles that are relevant for determining charges, (2)

communication of data for billing purposes, and (3) enforcement.[57] With conventional systems that use toll booths vehicle detection and payment are done manually and access is controlled by physical barriers. Electronic Toll Collection (ETC) systems perform the functions using various technologies. There are three types of ETC systems (Noordegraaf et al., 2009):

1. Roadside-only systems that use Automated Number Plate Recognition,
2. Dedicated Short Range Communications, and
3. In-vehicle-only systems that rely either on satellites or cellular networks.

Roadside-only systems and tag & beacon systems require roadside infrastructure and only record point data. To determine distance traveled a vehicle must be detected at a sequence of locations. (Distances can also be measured directly using odometers and, in the case of trucks, electronic tachographs.) In-vehicle-only systems track a vehicle's course and do not require roadside infrastructure although infrastructure-based technology may be used in tandem for enforcement purposes.

4.2. Component technologies

Each of the three types of ETC systems comprises one or more component technologies that each perform one or more of the three basic functions (Table 1).

Automated Number Plate Recognition

Automated Number Plate Recognition (ANPR) technology uses digital cameras and optical character recognition (OCR) software to record an image of a vehicle and its license plate. ANPR is used standalone with roadside-only systems although this requires collecting and processing images for every vehicle. Highway 407 in Toronto is an example of an open-road tolling scheme that uses ANPR to record vehicle usage for billing purposes. ANPR is more commonly used for enforcement because only violators have to be processed and — unlike other technologies — ANPR does not require that vehicles have equipment in working condition.

Table 1 : Congestion pricing functions and technologies

Technology	Road use measurement	Data communication	Enforcement
ANPR/ OCR	Location	Bills sent to user by post, deducted from bank account, etc.	√
DSRC	Location	√	√
GNSS (e.g. GPS)	Location	with GPRS	
Cellular networks	Location	with GPRS	
On Board Units	Location using GIS		

[57]The distinction between these functions is not clear cut (Bomberg et al., 2009), and additional functions can also be defined. Iseki et al. (2010) identify the three main functions as metering road usage, calculating charges, and communicating data to a collections agency. They bundle enforcement with data communication although enforcement is also a consideration for measuring road usage.

Smart cards		√	
Odometer/tachograph	Distance		
Dead reckoning	Distance		
Enforcement beacons			√
Enforcement transponders			√
Mobile monitors with readers			√

Notes: **ANPR** = Automatic Number Plate Recognition; **DSRC** = Dedicated Short Range Communications; **GIS** = Geographic Information System; **GNSS** = Global Navigation Satellite Systems; **GPRS =** General Packet Radio Service**; GPS** = Global Positioning System; **OCR** = Optical Character Recognition.

Sources: Various

Dedicated Short Range Communications

Dedicated Short Range Communications (DSRC) is a means of Automated Vehicle Identification (AVI) and is a member of the class of tag & beacon systems. Antennas mounted on overhead gantries communicate with tags or transponders on vehicles as they pass by. Like ANPR, DSRCtechnology can be used for all three basic functions: road usage measurement, data communication, and enforcement. It can also be used in conjunction with on-board units (see vehicle equipment below) to operate a zonal tolling system by activating a vehicle's on-board unit when it crosses into the zone, and deactivating it when the vehicle leaves the zone.

DSRC technologyoperates in the radio frequency or microwave range of the electromagnetic spectrum. Some tolling systems communicate in the infrared range. DSRC and infrared tolling systems differ in their susceptibility to interference[58] but have similar functionalities. Most tolled facilities in the US use DSRC technology[59] with E-ZPass being the most widely used system. Furthermore, the governing body responsible for E-ZPass — the Inter-Agency Group (http://www.e-zpassiag.com) — is trying to establish DSRC as a standard. For these reasons the assessment of tag & beacon technologies in this review is limited to DSRC.

Satellite systems

Global Positioning System (GPS) technology, developed by the US military, is a member of a class of systems called Global Navigation Satellite Systems (GNSS) which include the Russian GLONASS system and the European Galileo system that is under development. GPS is used for navigation and other military and civilian functions. GPS can be used in conjunction with General Packet Radio Service (GPRS) which is a cellular data service for communications, and with Geographical Information Systems (GIS) stored on On-board units (see below) that translate latitude and longitude data into locations on a digitized road map. GPS receivers detect position to within about 15 meters and accuracy can be enhanced using Differential GPS (DGPS), the European Geostationary Navigation

[58]DSRC systems are vulnerable to radio frequency interference while infrared systems can be disrupted by sunlight, heavy rain, dust and fog (Virginia Department of Transportation, 2007).

[59]US DOT (2009) provides lists of tolled bridges, tunnels and roads in the US as of January 1, 2009. A few entries in the lists do not clearly identify either the tolling technology that is used or whether the facility was operational. After deleting these entries, 340 facilities remain. Of these 340 facilities, 262 (77 percent) employ some form of ETC. Only four of them use infrared laser.

Overlay Service, the Wide Area Augmentation System (in North America), and the Multi-functional Satellite Augmentation System (in Asia) to 1-5 meters.[60]

A drawback of GPS is that satellite signals can be distorted by atmospheric aberrations, lost in tunnels, reflected by tall objects (signal multipath) and intercepted by overpasses and high buildings (the urban canyon effect). For backup, odometers can be used to record distance, and dead-reckoning can be used to keep track of location (although accuracy declines with distance traveled). The Galileo system is designed to be more accurate than GPS, but the project is several years behind schedule. The European GINA project is undertaking trials to assess the technical feasibility and economic viability of the Galileo system, and its application to Value Added Services.[61]

Cellular networks

Cellular networks are used by cellular phones. The most popular standard is Global System for Mobile (GSM) communications which uses Short Message Service (SMS). Cellular networks are within range of much of the roadway network and show promise as a means of road pricing although the application is not as well developed as it is for GPS. Like GPS, cellular networks do not require roadside infrastructure, and communications is possible anywhere rather than being restricted to gantries or locations where transponders have been installed.Cellular networks are also less susceptible to the urban canyon effect (Bomberg et al., 2009).

Vehicle equipment

Several in-vehicle technologies can be used for road pricing. All vehicles are equipped with a Vehicle Identification Number (VIN) that provides information about vehicle class, year of manufacture, make, model and weight. Transponders are used for communication using DSRC; they are typically mounted on a vehicle's windshield and identify the vehicle as it passes a roadside reader. Transponders vary widely in their costs and range of communications (Iseki et al, 2010). On-board units are more elaborate devices with computational capabilities, memory storage, and an interface for communication with DSRC, GPS, or cellular networks. Some systems use smart cards that store credit for payments and can be inserted into and removed from the OBU. Cellular phones with positioning capability are now very widespread, and could be used in lieu of transponders and OBUs to perform road pricing functions (Bomberg et al., 2009).[62]

A central question in the design of satellite- and cellular-based road-pricing schemes is how to distribute the road use measurement and charge computation tasks between the vehicle and other parties. The choice has implications for several assessment criteria considered in Section 4.4 including: in-vehicle equipment costs, flexibility, the scope for providing additional services, and privacy. There are several possibilities ranging from the "thin client" approach with minimum OBU functions to the "thick client" approach in which the OBU performs most of the functions. The design is described in more detail in Section 4.4.

---

[60]See GINA (2010b), http://ec.europa.eu/enterprise/policies/satnav/egnos/index_en.htm [December 11, 2010] and http://www8.garmin.com/aboutGPS/ [November 21, 2010].

[61]http://www.gina-project.eu/ [November 21, 2010].

[62]A start-up company BancPass has launched a service called PToll in Austin, Texas that will allow drivers to pay tolls by mobile phone using automated billing in lieu of purchasing a transponder (http://www.tollroadsnews.com/node/4567 and http://www.bancpass.com/ [November 21, 2010]). To set up an account, customers must record their vehicle's license plate and identify their method of payment. Operator transaction costs are expected to be lower than for transponders or manual toll collection (http://www.tollroadsnews.com/node/4567).

4.3. Technologies used in existing road pricing schemes

This section provides an overview of a sample of road pricing schemes to illustrate the range of technologies and technology combinations that are either being used for congestion pricing or can be adapted to implement it. The list in Table 2 covers four categories: HOT lanes, area-based schemes, distance-based HGV schemes, and US studies of distance-based charges for passenger vehicles.

High Occupancy Toll (HOT) lanes

SR-91 in Orange County, California was the first HOT lane facility in the world. Tolls vary by time of day. I-15 in San Diego was the second facility and the first to adopt responsive tolls; it is currently being expanded to a 20-mile-long managed-lanes facility with multiple entry and exit points, a moveable barrier, and tolls that are based on distance traveled. I-394 was the second facility to adopt responsive pricing and the first to separate toll lanes from general-purpose lanes using only striping rather than barriers.[63] All three facilities use transponders for road use measurement and communications. They rely primarily on visual inspection to enforce occupancy requirements as do other High Occupancy Vehicle (HOV) and HOT facilities. However, they use different technologies to supplement visual inspection as well as to verify payment and intercept violators (see Table 2).[64] Enforcement beacons are mounted on gantries and flash if a working transponder is detected when a vehicle passes underneath. If the light does not flash, enforcement personnel located downstream assess visually whether the vehicle has the occupancy required to be exempt from paying a toll. Enforcement transponders are mounted in enforcement vehicles and allow officers following a vehicle under scrutiny to determine whether it has a valid account. Mobile enforcement readers perform the same function as enforcement transponders, but are more flexible since the enforcement vehicle can either travel beside a vehicle under examination or park on the road.

Area-based schemes

The Singapore, London, and Stockholm schemes are the only area-based schemes designed to control congestion. Singapore's ERP scheme uses DSRC technology for road use measurement whereas London and Stockholm use ANPR. Ken Livingstone was determined to implement a congestion charge during his first term as mayor of London and he opted for ANPR as a proven and low-risk technology despite its high infrastructure and operating costs. During the Stockholm trial in 2006, both ANPR and transponders were used and approximately half the transactions were processed by each mode. ANPR was able to capture a larger-than-expected fraction of license plate images and it worked well even in bad weather. The transponders added appreciably to the system expense because they cost about $30 each and had a limited battery life[65]. When the scheme became permanent in 2007, transponders were abandoned for general use although they are still used for vehicles that are exempt from payment.

European distance-based Heavy Goods Vehicle schemes

In Switzerland, Austria, Germany, the Czech Republic, and the Slovak Republic, HGVs pay tolls proportional to distance traveled on some, or all, major roads. None of the HGV schemes is designed for congestion pricing although the Austrian and German technologies permit some differentiation of tolls by time and location.

[63] Several other HOT lane facilities are operational, and a number of new ones are either being built or planned. Some will feature time-of-day tolls, and others responsive tolls; see http://ops.fhwa.dot.gov/tolling_pricing/value_pricing/projects/allprojects.htm.

[64] Descriptions of the enforcement technologies here are taken from Halvorson and Buckeye (2006).

[65] http://www.tollroadsnews.com/node/3046 [December 6, 2010].

The Swiss toll applies to HGVs over 3.5 metric tons gross vehicle weight and is paid on the whole 71,000 km national road network. It is differentiated by emissions class[66] but not by type of road or time of day. Distance is recorded using a digital tachograph and a smart card. The unit is activated by roadside DSRC transponders when a vehicle enters the country, and it is deactivated when the vehicle exits. Charges are paid by inserting the smart card into a roadside terminal (Cottingham et al., 2007).

In contrast to the Swiss system, HGV tolls in Austria are only charged on the 2,060 km primary road network and are not differentiated by emissions class. An on-board unit called a "Go Box" is used for communications. It uses DSRC microwave technology, is attached to the windscreen, and can be easily set to register the number of axles on the truck and trailer (PMA/Tools Division AG, 2009).The Swiss on-board unit can be used in Austria as an alternative to the Go Box.[67]

The German HGV scheme Toll Collect applies to federal motorways and some secondary roads (12,000 km in total). Toll differentiation is similar to that in Switzerland, but the technology is more advanced in using GPS to measure distance and GSM communications. DSRC beacons are used for backup location information (Cottingham et al., 2007). The system is scalable in that more roads can be added, and the technology allows tolls to be differentiated by road type and time of day.

The Czech Republic and the Slovak Republic have the newest schemes. In the Czech Republic tolls are collected on 2,070 km of motorways and dual carriageways (PMA/Tools Division AG, 2009). DSRC technology is used; it is allegedly expensive to operate — probably in part because of an opaque tendering process (Schindler, 2007). The Slovak Republic has a satellite system. It differs from Germany's Toll Collect in that OBUs are obligatory. The OBUs are designed for easy installment to meet EC non-discriminatory regulations (Slovak Republic, 2010). Tolls are levied on 2,026 km of motorways, expressways, and first class roads.

Belgium, Denmark, France, Hungary, Slovenia, and Sweden are also considering HGV charging schemes that vary according to class of road covered, scope of toll differentiation, and technology (Noordegraaf et al., 2009; Dutch Ministry, 2009; GINA, 2010a). The UK Department for Transport (2004) proposed a national road pricing scheme for Great Britain with charging for HGVs to be implemented first, and other vehicles several years later. Tolls were to be differentiated by type of road, vehicle weight, number of axles, and emissions class, followed later by possible further differentiation by time of day and geographic area (Sorenson and Taylor, 2005). However, projected costs for the technology escalated, and the plan was eventually abandoned.

Table 2: Selected congestion pricing schemes and technologies

	Coverage and toll differentiation	Road use measurement	Data communications	Enforcement
HOT lanes				
SR-91/Orange County (1995)	TOD. HOV3+ exempt except eastbound on weekdays 4 pm -6 pm	FasTrak transponder	Transponder. Prepaid account	Cameras, enforcement beacons
I-394 (2005)	Responsive. HOV2+ exempt	MnPass transponder	Transponder. Prepaid account	Enforcement transponders, mobile readers
I-15 Managed lanes/San Diego (2009)	Responsive. Distance-based. HOV2+ exempt	FasTrak transponder	Transponder. Prepaid account	FasTrak smartcard identifies violators

[66]The toll rate is set to reflect the costs of health care, accidents, damage to buildings, and noise (Broaddus and Gertz, 2008).

[67]Most existing tolling schemes are not interoperable either between countries or within them.

Area-based schemes				
Singapore (1998)	Expressways & arterials + CBD + 1 cordon. TOD. By road and vehicle type	DSRC	DSRC and IVUs with smartcard	ANPR
London (2003)	Charging zone. Flat. Various exemptions	ANPR	Manual payment by various means	ANPR
Stockholm (2007)	Cordon. TOD. Various exemptions	ANPR. (Transponders for exempt vehicles.)	Monthly bill with payment by various means	ANPR
European distance-based heavy goods vehicle schemes				
Switzerland (2001)	All roads. Flat. By no. axles, emissions class, GVW > 3.5 tons	Tachograph and smartcard	DSRC and smartcard	ANPR & DSRC with GPS backup
Austria (2004)	Primary roads. Flat. By no. axles. GVW > 3.5 tons	DSRC using over 800 overhead gantries	OBUs “Go Box”	Cameras on selected gantries
Germany (2005)	Federal motorways & some secondary roads. Flat. By no. axles, emissions class. GVW > 12 tons	GPS	OBUs and GSM	Cameras, Mobile monitors DSRC backup using gantries
Czech Republic (2007)	Motorways and dual carriageways. By no. axles, emissions class. GVW > 3.5 tons	DSRC	OBUs"premid Box"	DSRC (microwave)
Slovak Republic (2010)	All motorways, expressways & some 1st class roads. By no. axles, emissions class, road type. GVW > 3.5 tons including buses	GPS	OBUs and GSM. Pre-pay or post-pay procedures.	Fixed & portable roadside installations & mobile monitors
US studies of distance-based charges for passenger vehicles				
Oregon (2004-2006)	Zonal. TOD	GPS & odometer (no GIS)	OBU and DSRC at gasoline stations	N/A
Puget Sound Regional Council (2002-2008)	Freeways & major arterials. TOD. By road type	GPS with OBU equipped with GIS	Cellular	N/A
Iowa (2005-2010)	Regional. Flat. By vehicle class, & road type eventually	GPS with odometer and OBU equipped with GIS	Cellular	Odometer & dead reckoning backup. GPS validation

Notes: **ANPR** = Automatic Number Plate Recognition; **DSRC** = Dedicated Short Range Communications; **GIS** = Geographical Information System; **GPS** = Global Positioning System; **GSM** = Global System for Mobile communications; **GVW** = Gross Vehicle Weight; **HOV** = High Occupancy Vehicle; **IVU** = In Vehicle Unit; **OBU** = On Board Unit; **TOD** = Time-of-day

Sources: Cottingham et al. (2007), Nash et al. (2008), GINA (2009), Noordegraaf et al. (2009), Iseki et al. (2009)

Plans for distance-based charges for passenger vehicles

Several countries have studied distance-based charges for passenger vehicles. As just noted, a scheme was planned for Great Britain, but in 2007 it was shelved in the face of strong public opposition. In 2008, the Dutch Parliament approved a national system of distance-based user charges with the fee per kilometer differentiated by vehicle emissions class and time of day. The so-called Dutch Mobility Plan would have used satellite technology. A trial using OBU displays and instant feedback on charges found that a majority of drivers were willing to change their behavior to avoid rush-hour travel when presented with the right incentives (The Institution of Engineering and Technology, 2010). However, the Plan was derailed when the Dutch government fell in February 2010.

The US has been considering a Vehicle Miles Traveled (VMT) fee as a long-run alternative to fuel taxes as the primary funding mechanism for roads. Depending on the technology used, the fee could be varied by time, distance, and location to price congestion. Several US experiments with regional distance-based pricing have been conducted that provide evidence on the technological possibilities and challenges (see Table 2). The Oregon Vehicle Miles Traveled Pricing Pilot Project (2004-2006) was designed to test the viability of distance-based charges as a replacement for fuel taxes.[68] Charges were defined by zone and set higher during AM and PM peak periods. Test vehicles were equipped with GPS devices that recorded mileage, but only aggregate distance was recorded and vehicle movements could not be tracked. The distance-based charge was paid automatically when vehicles refueled at participating gasoline stations, and the state fuel tax was deducted from the bill. The study found that GPS technology was reliable and assured privacy protection.

The Puget Sound Regional Council conducted a six-year study (2002-2008) of driver responses to network-wide facility-based tolls.[69] Tolls were differentiated by road type (higher on freeways than on arterials) and time of day (substantially higher during AM and PM peaks).[70]Unlike in the Oregon project, Geographical Information Systems (GIS) were required in combination with GPS to record separately distances traveled on freeways and on arterials. Test results were used to assess the merits of several road pricing schemes ranging from HOT lanes to all freeways and major arterials.

A third study, launched in 2005 and administered by the University of Iowa, is conducting a feasibility assessment of GPS-based tolling technology as well as gauging drivers' responses and public attitudes towards it.[71] Several test sites are located across the country. As in the Puget Sound study, GIS is used in combination with GPS to record distances within the region, to compute charges on the vehicle, and to download updates to the database. Only aggregate charging data is transmitted from the vehicle. Unlike in the Oregon and Puget Sound studies, tolls are flat.

4.4. Choice of technology

Any congestion pricing system has to perform the three basic functions of road use measurement, communication of billing data, and enforcement as well as computation of charges. The best technology choice for each function depends, among other things, on the type of charging scheme and the degree of toll differentiation to be implemented. Assessments are made here for the four technologies evaluated by Noordegraaf et al. (2009), referred to here as *ANPR*, *DSRC*, *Satellite*, and *Cellular*. Table 3 includes the criteria in Table 2 of Noordegraaf et al. (2009), except for privacy which is discussed subsequently, and adds several other assessment criteria.[72]Noordegraaf et al. rank the

[68]See Whitty (2007, 2009).

[69]See Puget Sound Regional Council (2002), Whitty (2009) and www.psrc.org/projects/trafficchoices.

[70]The tolls were virtual in the sense that test volunteers did not actually incur out-of-pocket costs.

[71] See www.roaduserstudy.org and Kuhl (2009).

[72]Further criteria such as system reliability, ease of audit and ease of integration with existing payment methods will also be mentioned.

systems for applications to distance-based charging. All the criteria are relevant for tolling facilities, cordons, and zones as well, although the rankings vary.

Table 3: Technology comparisons for distance-based charging

	ANPR	***DSRC***	***Satellite***	***Cellular***
System type	Roadside-only	Tag & beacon	In-vehicle only	In-vehicle only
Road use measurement	ANPR	On Board Unit	GNSS	Cellular network
Data communication		DSRC	GPRS	GPRS
Assessment criteria				
Location accuracy	+	+ +	+ +	+
Roadside infrastructure costs	- -	- -	+ +	+ +
Vehicle equipment costs	+ +	-	-	-
Flexibility	- -	+	+ +	+ +
Scalability	- -	- -	+ +	+ +

Notes: **ANPR** = Automatic Number Plate Recognition; **DSRC** = Dedicated Short Range Communications; **GNSS** = Global Navigation Satellite Systems; **GPRS =** General Packet Radio Service
Source: Noordegraaf et al. (2009, Table 2)

Location accuracy

Location accuracy refers to accuracy in detecting and identifying vehicles, and recording where they are. *ANPR* and *DSRC* use infrastructure in the vicinity of roadways, and can identify location precisely if they receive a proper signal. Modern *DSRC* technology has a recognition accuracy of 99% or better. *ANPR* has somewhat lower accuracy that varies by facility and with conditions. It can fail in bad weather or when the camera view of a number plate is obscured by dirt or other vehicles. Readability of license plates varies by country, and US states also differ in the design of their number plates. Furthermore, to provide adequate lines of sight on multilane highways, cameras must be mounted overhead on gantries rather than beside the roadway.

Satellite systems provide nearly ubiquitous coverage. Until recently their resolution was inferior to infrastructure-based technology and they could fail to distinguish between closely-spaced roads. However, enhancements such as Differential GPS have improved resolution to less than a lane width, and to back up temporarily lost signals inertial systems can maintain position measurements to within a meter for periods of several minutes (GINA, 2010b). GPS can also be used for area-based schemes by combining distance measurement with the detection of geobjects (GINA, 2010b). Nevertheless, as noted earlier GPS signals can still be disrupted by the urban canyon effect which is a greater problem in cities where accuracy is most important (Samuel, 2009). Another limitation is that commercial GIS maps do not always provide a consistent level of accuracy (Donath et al., 2009). Unlike GPS, *cellular* systems are not susceptible to the urban canyon effect. But their spatial resolution is limited by cellular tower density, which makes them more suitable for zonal than facility-based schemes. According to Bomberg et al. (2009), the resolution is probably sufficient to identify county boundaries, but not closely-spaced individual roads and perhaps not boundaries between smaller zones such as municipalities.

Roadside infrastructure and operating costs

ANPR and *DSRC* require roadside infrastructure whereas *Satellite* and *Cellular* do not unless *ANPR* and *DSRC* is used for enforcement. Roadside infrastructure is expensive to install, occupies space, is costly or impossible to relocate, requires maintenance, and is susceptible to vandalism. Given the high costs, *ANPR* and *DSRC* are likely to be economic only for tolling heavily-used facilities. Collection costs for existing facility-based systems in the US amount to roughly 16% of toll revenues (NSTIFC, 2009). Table 4 provides US cost data for toll plazas and toll administration. For a toll plaza, annual capital costs range from $27,000 to $44,000 (2009 dollars) and operating costs from $500 to $1,100. For toll administration, annual capital costs range from $44,000 to $86,000 and annual operation costs from $4,200 to $8,300. Since the technology is established, and scale economies have been largely exploited, these costs are not likely to evolve as rapidly as for *Satellite* or *Cellular*. Obtaining up-to-date information is difficult because many system providers do not divulge it for competitive reasons (Hamilton and Eliasson, 2010).

Infrastructure and operating costs for area-based urban schemes are more difficult to estimate since existing schemes are few, and highly varied. Estimated collection costs as a fraction of toll revenues are: 21% in Singapore, 22% for the Stockholm Trial, and 50-60% for London.[73,74] For the proposed cordon in San Francisco high-level estimates are $45 million for operating costs,$20 million for capital amortization, and $145 million for revenues after discounts (San Francisco County Transportation Authority, 2010) implying a cost-to-revenue ratio of 45%.

Table 4: Toll plaza and Toll administration costs

Unit Cost Element	**Lifetime** [Years]	**Capital Cost** $K, ($2009) (Source Year)	**O&M Cost** $K/year, ($2009) (Source Year)	**Description**
Toll plaza				
Electronic Toll Reader	10	2-4 (2001)	0.2 - 0.4 (2001)	Readers (per lane). O&M is estimated at 10% of capital cost.
High-Speed Camera	10	6-8 (2003)	0.3 - 0.7 (1995)	Cost includes one camera per two lanes.

[73] These figures are reported, along with sources, in Lindsey (2007, Table A1).

[74] The high costs for London have attracted much attention. According to the Fourth annual monitoring report (the last before the Western Extension) total annual scheme costs were £110 million (Transport for London, 2006, Table 9.1). Of this total, £20 million was spent on extra buses, £5 million on Transport for London administration, and the remaining £85 million for contractors. The principal contractor, Capita, was responsible for processing payments and fines, and employed subcontractors for much of the infrastructure (http://www.roadtraffic-technology.com/projects-/congestion/ [December 11, 2010]). A network of 331 camera sites is used to monitor traffic (http://www.tfl.gov.uk/assets/downloads/CC-Cameras.pdf [December 11, 2010]). The cameras require maintenance, andhuman input is required to read camera images that cannot be interpreted by the ANPR software. Costs are also incurred to administer the various means of payment: online, by SMS text messaging, by phone, automated telephone service, at shops, by post, and (as of January, 2011) by pre-registration (http://www.tfl.gov.uk/roadusers/congestioncharging/6744.aspx [December 12, 2010]).

Electronic Toll Collection Software	10	5-10 (1995)		Includes COTS software and database.
Electronic Toll Collection Structure	20	14 - 22 (1995)		Mainline structure.
Toll administration				
Toll Administration Hardware	5	4.3 - 6.4 (2004)	0.22 - 0.32 (2004)	Includes two workstations, printer, and modem. O&M estimated at 5% of capital costs.
Toll Administration Software	10	40 - 80 (1995)	4.0 - 8.0 (1995)	Includes local database and national database coordination. Software is COTS.

Notes: O&M = Operation and maintenance
Source: http://www.itscosts.its.dot.gov/its/benecost.nsf/SubsystemCostsAdjusted? OpenForm&Subsystem=-Toll+Plaza+(TP) [November 17, 2010]

Compared to area-based urban schemes, operating costs as a fraction of revenues are lower for the HGV schemes in Europe: 4% for Switzerland, 9% for Austria, and 16% for Germany (Broaddus and Gertz, 2008).[75] These lower percentages are attributable in part to the relatively high per-kilometer fees that trucks pay and the long distances they travel. The costs of satellite-based regional or national schemes that cover all vehicles are even harder to estimate — especially since the costs are sensitive to details of the technology choice (Glaister and Graham, 2008). The capital cost alone for the US is estimated to be of order $10 billion (NSTIFC, 2009).[76] The Dutch government had set a goal to limit administrative costs for the Dutch Mobility Plan to 5% of revenues. The GINA project has not yet developed a cost estimate or considered how enforcement will be implemented (GINA, 2010b). Since much of the costs of *Satellite* and *Cellular* systems are fixed, average total costs are likely to be lower for large countries. A further consideration in choosing a congestion pricing technology is that a system may be capable of providing additional services such as pricing of parking and insurance, navigation assistance, and so on. If so, the full system costs are not wholly attributable to the congestion pricing function. The attributable portion depends, among other things, on which service is considered incremental and this may be unclear.

Charge computation and in-vehicle equipment costs

ANPR does not require vehicle equipment except for readable license plates. The other three technologies require one or more types of equipment. *DSRC* systems require a transponder. Transponders can be active or passive. Passive transponders respond to roadside readers and use the energy from the incoming signal to reply. A Neology 6C tag is estimated to cost about $1.80 when mass produced.[77] Active transponders have their own power supply and can initiate data exchanges. A Kapsch 5.9Ghz transponder costs $20 to $30.[78]*Satellite* systems require an antenna, a power source, and (for systems that use GIS) digital maps with sufficient accuracy to locate a vehicle on a

[75] Capital costs are 4% of revenues for Switzerland, 3% for Austria, and 7% for Germany (Broaddus and Gertz, 2008).

[76] By comparison, a back-of-the-envelope calculation by Peter Samuel, editor of TollRoadsNews, suggests that a workable system using DSRC would cost of order $1 billion without coverage of rural collectors, and $1.5 billion with them (http://www.tollroadsnews.com/node/4409 [December 12, 2010]).

[77]http://www.tollroadsnews.com/node/4984 [November 18, 2010].

[78]http://www.tollroadsnews.com/node/4968[November 19, 2010].

particular road (Donath et al., 2009). OBUs for HGVs cost roughly €200 ($316) (Slovak Republic, 2010). For passenger vehicles the cost of a full-function OBU is approximately €100 ($158) (Hamilton and Eliasson, 2010). Mobile phones can be used as OBUs with *Cellular* systems, and this reduces the costs since most drivers already have a mobile phone although phones must somehow be linked to a given vehicle (Cottingham et al., 2007).[79]

As noted in Section 4.2, the tasks of road use measurement and charge computation can be distributed between the OBU unit and other parties in several ways.[80] The so-called "thin client" modeis one extreme in which the OBU sends all data to a control center which calculates the charge due for each road segment. At the opposite extreme is the "thick client", "fat client", or "smart client" modein which the OBU calculates the charges due for each segment, and transmits to the control center only the total amount due either for a trip or within an accounting period. The control center needs only to verify that the OBU is working properly.

Compared to the thin client mode,the thick client modeassures greater privacy protection and data security because detailed travel information does not leave a vehicle. Furthermore, if a breach of privacy does occur it is limited to a single vehicle rather than potentially widespread. The thick client modealso offers better system reliability because a memory failure or calculation error is only consequential for one vehicle. Furthermore, a vehicle operating in thick client modecan continue to pay tolls in the event of a temporary communication problem with the control center.

The thick client modealso has some disadvantages relative to the thin client mode. The OBU needs more memory, and greater computation and backup capabilities. The thick client is less flexible because changes to the network map and rate schedule need to be uploaded to each vehicle. Auditing is more difficult for users since charges are computed by the OBU, and system operators may be unable to break down the bill or explain how charges were assessed.[81]

Several approaches intermediate between the thin client and thick client modesare possible. One is the "distributed-role" or " third-party" approach which involves two centers or parties. One center receives the detailed data from the vehicle and computes the total charges due without knowing the identity of the driver. The first center sends the total to the second center which then arranges for the driver to be billed. Another intermediate approach, called "Anonymous Loop-Back Proxy Data Transmission", keepsdetailed travel information aboard the vehicle, but does the data processing and charge computation off the vehicle. The charges are then transmitted back to the vehicle which forwards them to a billing center. These and other intermediate approaches can be designed to exploit the advantages of the thin-client and thick-client modes while trying to avoid their disadvantages.[82]

[79]Cellular networks can, of course, be used for cell phone communications. However, Bomberg et al. (2009) point out that if every data transmission is treated as a phone call, operating costs are likely to be $5/month per vehicle or more (which compares unfavorably with the fuel tax as far as collecting revenues to pay for roads at low cost).

[80] The summary here draws heavily on International Working Group (2009) and Bomberg et al. (2009).

[81]As Bomberg et al. (2009) note, this drawback can be alleviated by retaining travel data on the OBU for a certain period so that it can be downloaded if necessary for auditing purposes.

[82]To maintain interoperability with existing toll collection schemes it is desirable that thick client and hybrid modes be able to operate in thin client mode (Bomberg et al., 2009).

Flexibility

System flexibility has several dimensions: to redeploy or expand the charging area; to modify or extend toll differentiation by road, time of day and vehicle characteristics; to add new technology and salvage old equipment; to add services such as route guidance, and so on. Infrastructure-based systems tend to be less flexible than in-vehicle systems in all respects. Since *ANPR* does not use in-vehicle equipment, *ANPR* systems cannot be changed or improved using vehicle technology. New cameras had to be installed when the Western extension was added in London, and it is unclear whether the cameras will have a salvage value after the Extension is abandoned. *DSRC* systems are less flexible than *Satellite* and *Cellular* systems because they tend to use OBUswith fewer capabilities (Noordegraaf et al., 2009) although (as noted above) the ease with which OBUs can be updated depends on whether a thin-client or thick-client mode is adopted.[83] With *Satellite* systems it is possible to modify cordon or zonal areas by redefining the related geo-objects (GINA, 2010b). Indeed, the German Toll Collect System was expanded to include secondary highways after HGVs began diverting onto them from the tolled motorways. And *Cellular* systems that operate across multiple jurisdictions, such as the one tested in the University of Iowa study, can be designed to accommodate changes in boundaries and charging policies (Iseki et al., 2010).

Scalability

The scalability of a technology is inversely related to the amount of roadside infrastructure it requires. *ANPR* and *DSRC* are at a disadvantage relative to *Satellite* and *Cellular* for regional, intercity, or national applications (NSTIFC, 2009; International Working Group, 2009; Grush, 2010).[84] Indeed, the London congestion charge could not economically be expanded to include Greater London even if the charging area remained as a single zone (Cottingham et al., 2007). *Cellular* systems based on text messaging have an advantage over other cellular technologies since they require less bandwidth and can be implemented using most cellular networks (Donath et al., 2009).

Privacy protection

Privacy protection has been a challenge for electronic road pricing at least since Hong Kong's electronic road pricing experiments in the 1980s. Privacy concerns could be assuaged by not keeping records of vehicle movements, but this would undermine system accountability. The use of *ANPR* in London and Stockholm for road use measurement has not evoked adverse reactions, and *ANPR* is used for enforcement in many other systems. A common misperception about *Satellite* systems is that they can be used to track the locations and movements of vehicles. In fact, vehicle receivers are passive and cannot transmit information back to satellites. However, both *Satellite* and *Cellular* systems have to transmit information to ground receivers, and they are more vulnerable to communications interception than *ANPR* or *DSRC* because information is transmitted over much longer distances. As noted earlier, this problem can be alleviated by adopting a thick-client mode and performing calculations with OBUs.[85]A further concern is that systems with OBUs that record travel

[83]In the University of Iowa experiment toll tables can be updated by downloading new tables over cellular networks (Kuhl, 2009).

[84]The infrastructure costs of *ANPR* and *DSRC* could be reduced considerably by locating readers only at high-volume nodes such as freeway interchanges (Cottingham et al., 2007; Bomberg et al., 2009). This would result in less-than-universal, but perhaps acceptable, coverage. Moreover, if a *Cellular* technology were available in tandem as part of a hybrid system, a vehicle that did not pass a reader within a given time period could use the *Cellular* option (Bomberg et al., 2009).

[85]The European Union has prohibited location information from leaving a vehicle and thus effectively legislated in favor of a thick-client mode. See Pecháčková (2008) for background to the legislation, and International Working Group (2009) for the *Sofia Memorandum* of March 2009 that established principles related to privacy protection.

speed could provide information that is used in court in case of an accident, or by insurance companies to adjust insurance rates (Sorensen and Taylor, 2005).

Enforcement

Enforcement is required to intercept vehicles with nonfunctional equipment or equipment that has been corrupted to provide false identification or other information. Of the four technologies, only *ANPR* does not rely on in-vehicle equipment and — except for problems with false or stolen number plates (as has occurred in London) — *ANPR* is robust to tampering. *ANPR* is the most common means of enforcement for facility-based and area-based congestion pricing schemes (cf. Table 2). Its big drawback for regional, intercity, or national applications is the high cost of establishing cameras throughout the network.

An alternative to cameras or other fixed infrastructure enforcement technologies are mobile monitors or readers such as those used by the German HGV system and on I-394 (cf. §4.3). Mobile enforcement is more cost-effective relative to stationary enforcement methods when only certain categories of vehicles (e.g. trucks) are tolled.

An additional enforcement task in the case of HOT lanes is to verify that vehicles choosing not to pay a toll meet the minimum vehicle occupancy requirement. Visual inspection has proved to be unreliable and it is also costly in labor input as well as land at facilities where an extra lane is built as an "enforcement zone" to facilitate observation. Research is underway on automated vehicle occupancy verification technologies that operate either on the roadside or inside vehicles. According to Poole (2009), most of the roadside systems cannot yet detect people in rear seats (which is required to enforce HOV3+ requirements) and in-vehicle systems face legal challenges due to privacy concerns.[86]

Scope for toll differentiation

Toll differentiation by vehicle characteristics such as number of axles, gross vehicle weight, and emissions class can be done using any of the technologies simply by registering the vehicle on a database (Noordegraaf et al., 2009). Differentiation by time of day is also straightforward except for ANPR. Differentiation by road type is done automatically with facility-based schemes. For distance-based schemes, DSRC is used with the Austrian HGV charge while GPS is used in Germany and the Slovak Republic.

Discounts and exemptions are provided to several categories of users in the London and Stockholm schemes by either registering them in a database or using DSRC (in Stockholm). Exemptions may also be desired for certain users on private roads, and for residents on local public roads who have paid for roads with property taxes. Such exemptions can be implemented by locating transponders or beacons at appropriate points on the network.

Additional services

The four technologies differ widely in their scope for providing services besides toll collection. For *ANPR* there is no scope other than using vehicle and license plate images for law enforcement. *DSRC* is also limited by the capabilities of OBUs and by the fact that communications is possible only when vehicles are near receivers. *Satellite* and *Cellular* systems offer greater potential for providing navigational aid and travel advisories, as well as charging for parking, insurance, and other

[86] As an alternative to a technology solution Poole suggests that policy be changed to require all vehicles using HOT lanes to carry a transponder and to require vehicles used for toll-free carpooling to pre-register. Consistent with this, the I-95 Express in Florida, which opened in December 2008, requires eligible vehicles (3+ passenger carpools, hybrids, and South Florida Vanpools) to pre-register in order to use the HOT lanes for free (http://www.95express.com/home/registration.shtm [December 11, 2010]).

services.Grush and Roth (2008) describe a system — similar to that used for telecommunications — that would perform these functions with charges differentiated by time, distance, and place.

4.5. Traveler information[87]

Information about travel conditions helps individuals make optimal mode, departure time, route, and other travel choice decisions.[88] For decades, travel information has been available from traditional sources such as newspapers, television, commercial radio and Highway Advisory Radio, as well as field devices such as changeable message signs. Television can provide relatively up-to-date information, but neither television nor newspapers provides en-route information. Message signs provide relevant location information without requiring drivers to do more than look at the signs. But signs are limited in how much information they can convey, and they cannot provide pre-trip travel information. Signs are also costly to install and maintain.

Starting in the 1990s, new information sources have become available: traffic websites, cellular phones, smart phones, Personal Intelligent Travel Assistants[89], and (in the US) publicly-operated 511 phone systems. In addition to navigational assistance and other services, these technologies can provide travelers with information about tolls. In the case of flat or time-of-day tolls, rates can be posted on the internet and viewed or downloaded at home. Conveying responsive tolls is more challenging because tolls can change rapidly, and pre-trip information may be obsolete by the time the toll is paid. Portable devices or in-vehicle screens that display real-time tolls may be useful (Cottingham et al., 2007; Noordegraaf et al., 2009). Nevertheless, making complex trip decisions en-route, or even pre-trip, may be so taxing that travelers will prefer to delegate decisions to OBUs or central computers used by the information service provider or toll operator.[90] Travelers could program an on-board unit to select a route with the shortest distance, shortest expected travel time, or lowest expected generalized cost. Websites already do this. For example, Traffic.com allows users to specify a starting point and ending point.[91] Using real-time information derived from proprietary and external sources Traffic.com determines two routes: a direct route and a route with the shortest current travel time. Information is provided for each route on distance, drive time, speed limit, delay, and average speed. A "jam factor" (on a scale of 0-10) is also reported for major roads along the routes, as well as the number of incidents in progress and an indication whether congestion is "building", "holding", or "clearing". Motorists can receive traffic alerts by SMS, automated voice call, and e-mail. The service is provided free in 52 US cities.

---

[87] This section draws on NCHRP (2009b).

[88] System operators also require information to set optimal congestion tolls — whether this be information on average annual daily traffic flows to set flat tolls, or real-time information on weather and incidents to set responsive tolls. Information can be obtained from many sources including conventional traffic counters, loop detectors, wireless radar, helicopter patrols, updates from road construction and maintenance departments, and reports from motorists.

[89] See Chorus and Timmermans (2011).

[90] There is evidence that people become less sensitive to tolls when they pay electronically rather than manually. Using data from US publicly owned toll facilities in the US, Finkelstein (2009) finds that an increase in the penetration rate of ETC payment was associated with a decline in the elasticity of driving with respect to toll levels. One explanation is that electronic payment (especially if it is delayed) is less salient to people than payment by cash. Paradoxically, this could mean that driving could become more elastic (again) if people delegate their travel decisions to external agents that weight toll payment fully when choosing between travel options.

[91] http://www.traffic.com [December 11, 2010].

5. Concluding remarks

Congestion pricing is an idea with a long academic pedigree that has gained credence amongst practitioners and policymakers. Yet congestion pricing in practice is still only a limited patchwork of schemes. Facility-based schemes dominate in North America. Europe has a few urban area-based congestion pricing schemes and a few intercity distance-based schemes for HGVs that are designed more for revenue generation and internalization of environmental costs than to control congestion. Several technologies are employed. The London and Stockholm designs use Automated Number Plate Recognition (ANPR). Singapore's Electronic Road Pricing, most toll roads, HOT lanes in the US, and three heavy goods vehicles schemes in Europe use Dedicated Short Range Communications (DSRC). The German and Slovak Republic HGV systems use satellite technology. Cellular technology has not yet been implemented anywhere although it is promising.

Road pricing technology should be chosen to best meet objectives. In addition to congestion relief, road pricing can be used to internalize the costs of emissions, accidents, noise, and road damage. It can also be used to pay for parking, to generate revenues, and to implement the beneficiary principle that the costs of roads should be paid by those who use them. Pricing congestion efficiently is arguably the most demanding goal in terms of differentiation by vehicle characteristics, location, time of day, and real-time driving conditions. This suggests that congestion pricing should drive the technology choice. But the economics of congestion pricing are more attractive if the technology that is chosen can perform other functions.

In this review we have described and evaluated four road-pricing technologies dubbed ANPR, DSRC, Satellite and Cellular. One general conclusion is that ANPR and DSRC are better suited for tolling individual facilities and urban areas where congestion is severe. Satellite and Cellular technologies appear to be more economical for pricing at larger geographical scales. Satellite systems are already used to differentiate by type of road, number of axles, and emissions class, and they are capable of differentiating by time of day as well.

With a nod to their locations relative to the ground, we will refer to the two pairs of technologies as "low-tech" and "high-tech" options. An overarching question for any application of road pricing is whether to pursue the low-tech option or the high-tech option.[92] Arguments can be made for each option on the basis of various assessment criteria. Simplicity favors the low-tech option. The history of successes and failures with road pricing suggests that simple systems that can be expanded and upgraded stand a better chance of successful implementation than systems that try to achieve theoretical perfection. Hong Kong and Cambridge experimented with sophisticated technologies and their complexity contributed to the failure of the plans to advance beyond the trial stage (Ison and Rye, 2005). Problems with the complex technology of the German Toll Collect HGV charge also forced the initial rollout to be aborted. Incompatibilities of the component technologies were partly responsible. As Noordegraaf et al. (2009) note, technologies can interact in unforeseen ways. Component technologies can be proven when tested in isolation, and yet may not work when assembled into a system. That said, the Toll Collect system has functioned smoothly since 2005. The SkyToll system in the Slovak Republic has as well, and trials with satellite technology in the US and the Netherlands have shown favorable results. Cellular technology requires further testing.

Interoperability is a problem for the low-tech option. The European Union has adopted a directive on interoperability of the European Electronic Toll Service whereby users must be able to use tolled facilities throughout the EU with just one subscription contract.[93] In the US, EZ-Pass dominates in the

[92]This is part of a more general questionof how to phase in efficient transportationpricing (e.g. Project MC-ICAM, nd).

[93]See http://ec.europa.eu/transport/its/studies/eets_en.htm [December 13, 2010]. Providers can use adapters to comply with the Directive. Yet doing so is costly, and Hamilton and Eliasson (2010) conclude that the costs of the Directive outweigh the benefits to users.

Northeastern and Midwestern states while FasTrak dominates in California. Incompatibility of the two systems may become a problem as tolling becomes more widespread. In contrast, satellite signals are non-proprietary and free of charge which makes travel between regions and countries easier as well as agreements on how to share revenues (Grush, 2010).

Most studies argue that congestion pricing should be implemented in stages rather than as a "big bang" so that it can be proven without committing huge resources or exposing large numbers of unwilling participants to potentially significant losses (Verhoef et al., 2007). Staged implementation appears to favor the low-tech option. Nevertheless, Grush (2010) argues that once a satellite-based system has been established in one region, additional regions can be added incrementally without major effort. A high-tech option can also be introduced as an alternative to a low-tech option on a voluntary basis (Bomberg et al., 2009), possibly using incentives such as value-added services rather than mandates (Iseki et al., 2010).

For intercity transport the high-tech option appears to dominate the low-tech option, but so far only two high-tech schemes for HGVs exist. There are several reasons to implement distance-based tolls for HGVs first. Many HGVs are already equipped with GPS-based fleet management systems, and additional equipment for levying tolls can be added at moderate cost. Toll collection costs per unit of revenue are likely to be much lower than for passenger vehicles because trucks are driven long distances and pay higher tolls per kilometer. Truck drivers and shippers are also already familiar with the technology, and seem less concerned about privacy than automobile drivers. Some new automobiles are equipped with on-board units that make distance-based pricing practical at reasonable cost.[94] However, older vehicles lack this equipment and retrofitting is difficult and costly (Whitty, 2009). Studies also show that people have been very slow to adopt low-tech Electronic Tolling Technology (Amromin et al., 2007; Finkelstein, 2009) and the same could prove to be true for high-tech options. A prolonged phase-in period for passenger vehicles is therefore likely before satellite (or possibly cellular) technology becomes the standard.

Universal coverage of the road network is also unlikely to be achieved for some time — in part because toll enforcement is not yet possible without some form of roadside infrastructure that is expensive to build and operate on a wide geographical scale. Traffic diversion from tolled to untolled roads is therefore a potential problem, and the challenges of second-best pricing discussed in Sections 2 and 3 will remain relevant.

As noted earlier, several studies have evaluated the relative merits of road pricing technologies. Ukkusuri et al. (2007) evaluate six candidate road pricing technologies[95] on the basis of sixteen criteria using a multi-criteria decision-making algorithm. Scores for the criteria were imputed from the PRoGReSS project reports (European Commission, 2004). Radio Frequency Identification gets the highest ranking, followed by ANPR. Ukkusuri et al. conclude (p.8) that "GPS systems are not necessary for zone-based pricing, since they are required only for continuous monitoring. Other, less costly technologies like ANPR and DSRC can be used."

Two studies mentioned earlier evaluate technology choices for a mileage-based user fee. Bomberg et al. (2009) assess a fee for Texas. Their view is nuanced, but they appear to favor a two-track approach combining a low-tech option with a relatively cheap high-tech option that does not rely on value-added services to make it worthwhile. NCHRP (2009a) implicitly favors *Satellite* or *Cellular* as a

[94]For example, since 1996 vehicles sold in the UShave been equipped with a data bus called an on-Board Data link 2 that — in conjunction with an OBU — can be used to compute distance and communicate with a back office via SMS text messaging (Donath et al., 2009).

[95]The technologies are: manual toll booths, ANPR, DSRC, GPS, infrared communications, and Radio Frequency Identification (RFID). Ukkusuri et al. define DSRC as a subset of RFID that employs 5.9 GHz communications.

replacement for the fuel tax. The authors consider[96] DSRC too costly for comprehensive tolling of a road network for the purpose of revenue generation although they view DSRC as a candidate for congestion pricing.

A fourth study by Iseki et al. (2010) assesses road pricing technologies for various purposes. It concludes that the optimal choice of technology depend on the objectives, with the most important factors that determine the best choice being the geographical scale of the road network that is tolled and the complexity of the pricing scheme. For revenue generation at the regional or state level they lean towards GNSS/GSM technology.

Given the many arguments for and against low-tech and high-tech options, and the importance of ancillary concerns such as public acceptability, we hesitate to place bets. It appears that both low-tech and high-tech options will be pursued in the near term. Given their advantages in terms of scale economies, value-added services, and revenue generation as a supplement or replacement for fuel taxes, it seems plausible that either Satellite or Cellular technologies will come into widespread use in the longer term. If so, it makes sense to use them for congestion pricing as well as other functions.

Acknowledgments

We are indebted to SatishUkkusuri, Gillian Schaeffer and three anonymous referees for many helpful comments and corrections.

Role of the funding source

De Palma would like to thank the InstitutUniversitaire de France. Lindsey gratefully acknowledges financial support from the Social Sciences and Humanities Research Council of Canada. The SSHRC was not involved in any specific aspects of this review or in the decision to submit the paper for publication.

References

Amromin, G., C. Jankowski and R.D. Porter (2007), "Transforming payment choices by doubling fees on the Illinois Tollway", *Economic Perspectives 2Q,* 22-47.

Anderson, S. and R. Renault (2011), "Price discrimination", in de Palma, A., R. Lindsey, E. Quinet and R. Vickerman, eds., *Handbook in Transport Economics,* Edward Elgar.

Arnott, R. (2011), "Parking economics", in de Palma, A., R. Lindsey, E. Quinet and R. Vickerman, eds., *Handbook in Transport Economics*, Edward Elgar.

Arnott, R. and M. Kraus (1998), "When are anonymous congestion charges consistent with marginal cost pricing?", *Journal of Public Economics 67(1),* 45-64.

Arnott, R., T. Rave and R. Schöb (2005), *Alleviating Urban Traffic Congestion,* Cambridge: MIT Press.

Bansal, M. and D. Smith (2010), Final Report, *"CLRP Aspirations" Scenario*, TPB Scenario, National Capital Region Transportation Planning Board, Study Metropolitan Washington Council of Governments, September 8 (http://www.mwcog.org/uploads/committee-documents/-ZV5YWVhb20100909154020.pdf [December 7, 2010]).

Bonsall, P. and J. Knockaert (2008), *Recommendations for differentiated charges for car drivers*, Deliverable 9.3, DIFFERENT: User Reaction and Efficient Differentiation of Charges and Tolls (www.different-project.eu [July 29, 2009]).

[96] See Section 7.9.

Bonsall, P., J. Shires, J., Maule, B., Matthews and J. Beale (2007), "Responses to complex pricing signals: Theory, evidence and implications for road pricing", *Transportation Research Part A 41(7)*, 672-683.

Bordoff, J.E. and P.J. Noel (2008), "Pay-As-You-Drive auto insurance: A simple way to reduce driving-related harms and increase equity", The Hamilton Project, the Brookings Institution (www.brookings.edu/~/media/Files/rc/papers/2008/07_payd_bordoffnoel/07_payd_bordoffnoel.pdf [April 1, 2009]).

Boyce, D. (2007), "Future research on urban transportation network modeling", *Regional Science and Urban Economics 37(4)*, 472-481.

Braid, R.M. (1996), "Peak-load pricing of a transportation route with an unpriced substitute", *Journal of Urban Economics 40,* 179-197.

Broaddus, A. and C. Gertz (2008), *Tolling heavy goods vehicles: An overview of European practice and lessons from German experience*, January 31 (http://www.trb-pricing.org/index.php?option=com_docman&task=doc_download&gid=8&Itemid=91, [August 22, 2009]).

Calfee, J. and C. Winston (1998), "The value of automobile travel time: Implications for congestion policy", *Journal of Public Economics 69,* 83-102.

Carey, M. and A. Srinivasan (1993), "Externalities, average and marginal costs, and tolls on congested networks with time-varying flows", *Operations Research 41(1),* 217-231.

Chew, V. (2008), "Electronic road pricing: Developments after phase I", National Library Singapore, Singapore Infopedia (http://infopedia.nl.sg/articles/SIP_1386_2009-01-05.html [July 25, 2009]).

Chorus, C.G. and H.J.P. Timmermans (2011), "Personal Intelligent Travel Assistants", in de Palma, A., R. Lindsey, E. Quinet and R. Vickerman, eds., *Handbook in Transport Economics,* Edward Elgar.

Conway, A. and M.C. Walton (2009), "Policy options for truck user charging", paper presented at the 88th Transportation Research Board Annual Meeting 2009 Conference, CD Paper #09-2699, Washington, D.C.

Cottingham, D.N., A.R. Beresford and R.K. Harle (2007), "Survey of technologies for the implementation of national-scale road user charging", *Transport Reviews 27(4)*, 499-523.

Daganzo, C.F. (1994), "The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic model", *Transportation Research Part B 28(4),* 269-287.

Daniel, J.I. and K. Bekka (2000), "The environmental impact of highway congestion pricing", *Journal of Urban Economics 47(2)*, 180-215.

De Palma, A. and M. Fosgerau (2011), "Dynamic and static congestion models", in de Palma, A., R. Lindsey, E. Quinet and R. Vickerman, eds., *Handbook in Transport Economics*, Edward Elgar.

De Palma, A., M. Kilani and R. Lindsey (2005a), "Congestion pricing on an urban road network: A study using the dynamic traffic simulator METROPOLIS", *Transportation Research Part A 39(7)*, 588-611.

De Palma, A., M. Kilani and R. Lindsey (2005b), "A comparison of second-best and third-best tolling schemes on a road network", *Transportation Research Record 1932*, 89-96.

De Palma, A., R. Lindsey and S. Proost, eds., (2007), *Investment and the Use of Tax and Toll Revenues in the Transport Sector*, Research in Transportation Economics, Vol. 19, Amsterdam: Elsevier.

De Palma, A., R. Lindsey and E. Quinet (2004) “Time-varying road pricing and choice of toll locations”, in Santos, G. (ed.), *Road Pricing: Theory and Evidence*, Research in Transportation Economics, Vol. 9, Elsevier Science, 107-131.

Donath, M., A. Gorjestani, C. Shankwitz, R. Hoglund, E. Arpin, P-M. Cheng, A. Menon and B. Newstrom (2009), “Technology enabling near-term nationwide implementation of distance based road user fees,” Intelligent Transportation Systems Institute, Center for Transportation Studies, University of Minnesota, June 2009, Report no. CTS 09-20 (www.its.umn.edu /Publications/ResearchReports/reportdetail.html?id=1790 [July 16, 2009]).

Dong, J., S. Erdogan, C.-C. Lu and H. S. Mahmassani (2007), “State-dependent pricing for real-time freeway management: Static, reactive and anticipatory”, paper presented at the 86th Annual Meeting of the Transportation Research Board, Washington, D.C., Conference CD Paper No. 07-2109.

Dutch Ministry of Transport, Public Works and Water (2009), Management DPFM on Road Pricing, Connekt Information Meeting, June 30.

Ecola, L. and T. Light (2009), Equity and congestion pricing: A review of the evidence. Technical Report, Rand Transportation, Space, and Technology (http://www.rand.org/pubs /technical_reports/TR680/ [June 5, 2009]).

Eichler, M. D., G. K. Miller and J. Park (2008), Evaluating alternative scenarios for a network of variably priced highway lanes in the Metropolitan Washington Region, Final Report, National Capital Region Transportation Planning Board, February (http://www.mwcog.org/uploads/committee-documents/aF5fWVlW20080314161420.pdf [December 7, 2010]).

Eliasson, J. (2009), “A cost-benefit analysis of the Stockholm congestion charging system”, *Transportation Research Part A 43(4),* 468-480.

Eliasson, J., L. Hultkrantz and L. SmidfeltRosqvist (2009a),“Introduction: Stockholm Congestion Charging Trial”, *Transportation Research Part A 43(3)*,237-239.

Eliasson, J., L. Hultkrantz, L. Nerhagen and L. SmidfeltRosqvist (2009b), “The Stockholm congestion-charging trial 2006: Overview of effects”, *Transportation Research Part A 43(3),* 240-250.

European Commission, Competitive and Sustainable Growth Programme (2004), *PRoGR€SS Project 2000-CM.10390, Deliverable 9, Final Report*, July (http://www.progress-project.org/Progress/pdf/Main%20Project%20Report.pdf [December 7, 2010]).

FHWA (2008), Transportation Value Pricing Projects in the United States (http://www.ecy.wa.gov/climatechange/2008CATdocs/IWG/tran/082208_value_pricing_projects_in_us_for_t3.pdf [August 31, 2009]).

Finkelstein, A. (2009), “E-ZTAX: Tax salience and tax rates”, *The Quarterly Journal of Economics 124(3)*, 969-1010.

Friesz, T., C. Kwon and D. Bernstein (2008), “Analytical dynamic traffic assignment models”, in Hensher, D.A. and K.J. Button (eds.), *Handbook of Transport Modelling, Vol. 1*, 2nd ed., Oxford: Elsevier Science, 221-237.

Ghali, M.O. and M.J. Smith (1995), “A model for the dynamic system optimum traffic assignment problem”, *Transportation Research Part B 29(3)*, 155-170.

GINA (GNSS for INnovative road Applications) (2010a), “Some RUC schemes” (http://www.gina-project.eu/en/about-ruc/ruc-schemes/ [December 11, 2010]).

GINA (GNSS for INnovative road Applications) (2010b), “How can EGNOS and Galileo contribute to innovative road pricing policy? First findings and proposals from GINA project”, Brussels, 1

October 2010 (http://www.gina-project.eu/media/workshop/GINA_Workshop_Summary-.final.pdf [November 21, 2010]).

Glaister, S. and D.J. Graham (2005), "An evaluation of national road user charging in England", *Transportation Research Part A 39(7–9),* 632–650.

Glaister, S. and D.J. Graham (2008), "National road pricing in Great Britain: Is it fair and practical?", in Richardson, H. and C. Bae (eds.), *Road Congestion Pricing in Europe: Implications for the United States,* Cheltenham, UK; Northampton, MA: Edward Elgar, 57-97.

Gómez-Ibáñez, J.A. and J.R. Meyer (1993), *Going Private: The international Experience with Transport Privatization,* Washington, D.C.: The Brookings Institution.

Graham, D.J. (2007), "Variable returns to agglomeration and the effect of road traffic congestion", *Journal of Urban Economics 62,* 103-120.

Graham, D. and K. Van Dender (2008), "Pricing congestion with heterogeneous agglomeration externalities and workers", December, working paper (www.aeaweb.org/assa/2009/retrieve.php?pdfid=303 [December 7, 2010]).

Grush, B. (2010), "10 Reasons GNSS Tolling is Better than Microwave", Skymeter Corporation, modified from a paper submitted to the 10th Slovenian Road and Traffic Congress, October (www.skymetercorp.com/white_papers/10Reasons.pdf [December 5, 2010]).

Grush, B. and G. Roth (2008), "Paying for roads in the 21st Century with TDP Pricing", paper presented at the 88th Annual Meeting of the Transportation Research Board, Washington, D.C. Conference CD Paper No. 09-0222.

Halvorson, R. and K.R. Buckeye (2006), "High-Occupancy Toll lane innovations: I-394 MnPASS", *Public Works Management & Policy 10(3),* 242-255.

Hamilton, C.J. and J. Eliasson (2010), "A cost-benefit analysis of the European Directive on EFC interoperability - Interoperability pushed too far", working paper, Centre for Transport Studies, Royal Institute of Technology, August.

Hearn, D. W. and M. B. Yildirim (2002), "A toll pricing framework for traffic assignment problems with elastic demand", in Gendreau, M. and P. Marcotte (eds.), *Transportation and network analysis: Current trends,* Norwell: Kluwer Academic, 135-145.

Hensher, D.A. (2006), "Integrating accident and travel delay externalities in an urban speed reduction context", *Journal of Urban Economics 26(4),* 521-534.

Hussain, S. and N.A. Parker (2006), "Pavement damage and road pricing", paper presented at the 85th Annual Meeting of the Transportation Research Board. Conference CD Paper No. 06-1342.

International Working Group on Data Protection in Telecommunications (2009), Report and Guidance on Road Pricing - "Sofia Memorandum", Sofia (Bulgaria) 675.38.12 13, March 2009 (http://www.datenschutz-berlin.de/attachments/647/WP_Road_Pricing_Final_675.38.12.pdf?1264411301 [December 5, 2010]).

Iseki, H., B.D. Taylor and A. Demisch (2010), "Examining the linkages between electronic roadway tolling technologies and road pricing policy objectives", 89th Annual Meeting of the Transportation Research Board, Washington, D.C. Conference CD Paper No. 10-4033.

Ison, S. and T. Rye (2005), "Implementing road user charging: The lessons learnt from Hong Kong, Cambridge and Central London", *Transport Reviews 25(4),* 451-465.

Jensen-Butler, C., B. Sloth, M.M. Larsen, B. Madsen and O.A. Nielsen, eds. (2008), *Road Pricing, the Economy and the Environment,* Advances in Spatial Science, Springer Verlag.

Johansson-Stenman, O. and T. Sterner (1998), "What is the scope for environmental road pricing?", in Button, K.J. and E.T. Verhoef (eds.), *Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility*, Cheltenham, UK: Edward Elgar, 150-170.

Kuhl, J. (2009), Presentation at Stakeholder Perspectives Panel, 2009 Symposium on Mileage-Based User Fees, April 14-15, 2009, Austin, Texas, Symposium Proceedings (http://utcm.tamu.edu/mbuf/proceedings/ [July 16, 2009]).

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

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

Leape, J. (2006), "The London Congestion Charge", *Journal of Economic Perspectives 20(4),* 157-176.

Lindsey, R. (2006), "Do economists reach a conclusion on highway pricing?: The intellectual history of an idea", *Econ Journal Watch (http://www.econjournalwatch.org) 3(2)*, 292-379.

Lindsey, R. (2007), Congestion relief: Assessing the case for road tolls in Canada, C. D. Howe Institute Commentary 248 (www.cdhowe.org [May 2007]).

Lindsey, R. (2009), "State-dependent congestion pricing with reference-dependent preferences", University of Alberta working paper 2010-04.

Lindsey, R., V. van den Berg and E.T. Verhoef (2010), "Step by step: Revisiting step tolling in the bottleneck model", Tinbergen Institute Discussion Paper TI 2010-118/3 (http://www.tinbergen.nl/discussionpapers/10118.pdf)

Lindsey, R. and E.T. Verhoef (2001), "Traffic congestion and congestion pricing", in Button, K.J. and D.A. Hensher (eds.), *Handbook of Transport Systems and Traffic Control*, Oxford: Elsevier Science, 77-105.

Lou, Y., Y. Yin and J.A. Laval (2011), "Optimal dynamic pricing strategies for High-Occupancy/Toll lanes", *Transportation Research Part C 19(1)*, 64-74.

Mackie, P. (2005), "The London congestion charge: A tentative economic appraisal: A comment on the paper by Prud'homme and Bocajero", *Transport Policy 12,* 288-290.

Maruyama, T. and A. Sumalee (2007), "Efficiency and equity comparison of cordon- and area-based road pricing schemes using a trip-chain equilibrium model", *Transportation Research Part A 41(7)*, 655-671.

May, A., S. Shepherd, A. Sumalee and A. Koh (2008), "Design tools for road pricing cordons", in Richardson, H. and C. Bae (eds.), *Road Congestion Pricing in Europe: Implications for the United States*, Cheltenham, UK; Northampton, MA: Edward Elgar, pp. 138-155.

Nash, C. with contributions from partners (2003), Project UNITE (UNIfication of accounts and marginal costs for Transport Efficiency), Final Technical Report, Fifth Framework Competitive And Sustainable Growth (Growth) Programme, Commissioned by European Commission, DG TREN, www.its.leeds.ac.uk/UNITE.

National Cooperative Highway Research Program (2009a), Implementable Strategies for Shifting to Direct Usage-Based Charges for Transportation Funding, Project 20-24(69), Transportation Research Board, Washington, DC, June (http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_w143.pdf [December 12, 2010]).

National Cooperative Highway Research (2009b), NCHRPProgram Synthesis 399, Real-Time Traveler Information Systems, Transportation Research Board, Washington, DC (http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_syn_399.pdf [July 14, 2009]).

National Surface Transportation Infrastructure Financing Commission. 2009. Paying Our Way: A New Framework for Transportation Finance. www.financecommission.dot.gov/Documents/NSTIF_Commission_Final_Report_Advance%20Copy_Feb09.pdf [March 4, 2009].

Noordegraaf, D.V., B. Heijligers, O. van de Riet and B. van Wee (2009), "Technology options for distance-based road user charging schemes", paper presented at the 88th Annual Meeting of the Transportation Research Board, Washington, D.C. Conference CD Paper No. 09-2477.

Oldridge, B. (1995), "Congestion metering in Cambridge City, United Kingdom", in Johansson, B. and L-G Mattsson (eds.), *Road Pricing: Theory, Empirical Assessment and Policy,* Boston: Kluwer, 131-140.

Olszewski, P. and L. Xie (2005), "Modelling the effects of road pricing on traffic in Singapore", *Transportation Research Part A 39(7-9)*, 755-772.

Parry, I.W.H. (2005), "Is Pay-as-You-Drive insurance a better way to reduce gasoline than gasoline taxes?", *American Economic Review (Papers and Proceedings) 95(2)*, 188-293.

Parry, I.W.H. (2009), "Pricing urban congestion", *Annual Review of Resource Economics 1(1)*, 461-484.

Parry, I.W.H. and A. Bento (2001), "Revenue recycling and the welfare effects of road pricing", *Scandinavian Journal of Economics 103(4),* 645-671.

Pecháčková, H. (2008), "Toll collection & privacy and data protection issues", European Commission, Directorate-General Justice, Freedom and Security, Unit C5 — Data protection, FIA workshop on road pricing & traffic restrictions, February 6-7, Barcelona (http://www.racc.es/pub/ficheros/adjuntos/adjuntos_dataprotection_eu_jzq_7737694e.pdf [December 12, 2010]).

Peeta, S., W. Zhou and P. Zhang (2004), "Modeling and mitigating of car-truck interactions on freeways", *Transportation Research Record 1899,* 117-126.

Pigou, A.C. (1912), *Wealth and Welfare,* London: Macmillan.

PMA/Tools Division AG (2009), Toll systems and devices – An overview, May (http://www.pma-tools.net/de/Files/Mautsysteme%20und%20Geraete_EN.pdf [November 21, 2010]).

Poole, R. (2009), "Automating managed lanes enforcement", paper presented at the 88th Annual Meeting of the Transportation Research Board, Washington, D.C. Conference CD Paper No. 09-0385.

Poole, R.W., Jr. and C.K. Orski (2003), HOT Networks: A new plan for congestion relief and better transit, Reason Public Policy Institute Policy Study No. 305.

Project MC-ICAM (nd), Implementation of Marginal Cost Pricing in Transport – Integrated Conceptual and Applied Model Analysis (www.its.leeds.ac.uk/projects/mcicam). Funded by the European Commission, Contract No: GRD1/2000/25475-SI2.316057, 2001-02.

Proost, S. and K. Van Dender (1998), "Variabilization of car taxes and externalities", in Button, K.J. and E.T. Verhoef (eds.), *Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility*, Cheltenham, UK: Edward Elgar, 136-149.

Prud'homme, R. and J.P. Bocarejo (2005), "The London congestion charge: a tentative economic appraisal", *Transport Policy 12,* 279-287.

Prud'homme, R. and P. Kopp (2006), "The Stockholm toll: An economic evaluation", September 7 (http://www.sika-institute.se/Doclib/2006/pm_060907_prud.pdf [December 14, 2010]).

Puget Sound Regional Council (2002), Summary of the Puget Sound Regional Council's Examination of Transportation Pricing Strategies, Seattle, January (http://www.psrc.org/projects/pricing/summary.pdf, [September 3, 2009]).

Pustral UGM (2010), "Swedish city of Gothenburg plans to introduce congestion charging", The Center for Transportation and Logistics Studies, UniversitasGadjahMada, February 10, (http://www.pustral-ugm.org/en/?p=39 [November 25, 2010]).

Richardson, H. and C-H C. Bae, eds. (2008), *Road Congestion Pricing in Europe: Implications for the United States*, Edward Elgar: Cheltenham, UK and Northampton, MA.

Rotaris, L., R. Danielis, E. Marcucci and J. Massiani (2010), "The urban road pricing scheme to curb pollution in Milan, Italy: Description, impacts and preliminary cost–benefit analysis assessment",*Transportation Research Part A 44(5)*, 359-375.

Samuel, P. (2009), post to Congestion Pricing Forum listserv on June 17.

Samuel, P., R. Poole Jr. and J. Holguín-Veras (2002), Toll truckways: A new path toward safer and more efficient freight transportation. Reason Public Policy Institute (http://www.rppi.org).

San Francisco County Transportation Authority (2010), San Francisco Mobility, Access and Pricing Study, Draft Final Report, December (http://www.sfcta.org/images/stories/Executive/Meetings/cac/2010/12dec/MAPS-Enclosure.pdf [December 8, 2010]).

Santos, G. (2004), "Urban road pricing in the U.K.", in Santos, G. (ed.), Road Pricing: Theory and Evidence, Research in Transportation Economics, Vol. 9, Elsevier Science, 251-282.

Santos, G. (2008), "London congestion charging", in Burtless, G. and J. Rothenberg Pack (eds.), *Brookings Wharton Papers on Urban Affairs: 2008*, The Brookings Institution, 177-207.

Santos, G. and G. Fraser (2006), "Road pricing: Lessons from London", *Economic Policy* April, 265-310.

Schade, J. and B. Schlag (eds.) (2003), *Acceptability of Transport Pricing Strategies*, Amsterdam: Elsevier.

Schrank, D. and T. Lomax (2009), The 2009 Urban Mobility Report. College Station: Texas Transportation Institute, Texas A&M University. http://mobility.tamu.edu [July 16, 2009].

Shoup, D.C. (2005), *The High Cost of Free Parking,* Chicago, Illinois and Washington, D.C.: APA Planners Press.

Slovak Republic (2010), "Multi-lane free-flow electronic tolling in the Slovak Republic" (http://www.asecap.com/english/documents/Multi-Lane_Free-Flow_Electronic_Tolling_in-_the_Slo.pdf [March 24, 2010]).

Small, K.A. (2008), "Private provision of highways: Economic issues", working paper, February 5 (www.socsci.uci.edu/~ksmall/Private%20Hwy.pdf [March 7, 2009])

Small, K.A. and E.T. Verhoef (2007), *The Economics of Urban Transportation*, London: Routledge.

Small, K.A., C. Winston and J. Yan (2006), "Differentiated road pricing, express lanes, and carpools: Exploiting heterogeneous preferences in policy design", *Brookings-Wharton Papers on Urban Affairs,* 53-96.

Sorensen, P.A. and B.D. Taylor (2005), "Review and synthesis of road-use metering and charging systems", Transportation Research Board, Washington, DC (http://pubsindex.trb.org/default.asp [July 17, 2009]).

Steimetz, S. (2008), "Defensive driving and the external costs of accidents and travel delays", *Transportation Research Part B 42(9),*703-724.

Swan, P. and M. Belzer (2008), "Empirical evidence of toll road traffic diversion and implications for highway infrastructure privatization", paper presented at 87th Annual Meeting of the Transportation Research Board, Washington, D.C. Conference CD Paper No. 08-2727.

The Institution of Engineering and Technology (2010), "Road-pricing trial shows incentives can change driver behaviour", March 4 (http://kn.theiet.org/news/mar10/road-price-trial-results.cfm [December 11, 2010]).

Trannoy, A. (2011), "Equity dimensions of transport policies", in de Palma, A., R. Lindsey, E. Quinet and R. Vickerman, eds., *Handbook in Transport Economics,* Edward Elgar.

Transport and the Environment (2010), Understanding the effects of introducing lorry charging in Europe, July (www.transportenvironment.org/Publications/prep_hand_out/lid/592 [November 22, 2010]).

Transport for London (2006), Central London Congestion Charging: Impacts Monitoring, Fourth Annual Report, June (www.tfl.gov.uk/assets/downloads/FourthAnnualReportFinal.pdf [December 12, 2010]).

Transport for London (2007), Central London Congestion Charging: Impacts Monitoring, Fifth Annual Report, July (www.tfl.gov.uk/assets/downloads/fifth-annual-impacts-monitoring-report-2007-07-07.pdf [August 24, 2009]).

Transportation Research Board (2000), Highway Capacity Manual, Transportation Research Board, National Research Council, Washington D.C.

Tsekeris, T. and S. Voß (2008), "Design and evaluation of road pricing: state-of-the-art and methodological advances", Netnomics DOI 10.1007/s11066-008-9024-z.

UK Department for Transport (2004), Feasibility Study of Road Pricing in the UK Report, London (http://www.dft.gov.uk/pgr/roads/introtoroads/roadcongestion/feasibilitystudy/studyreport / [August 12, 2009])

Ukkusuri, S.V.S.K., A. Karoonsoontawong, S. T.Waller and K.M. Kockelman (2007), "Congestion pricing technologies: A comparative evaluation", Chapter 4 in Filip N. Gustavsson (ed.), *New Transportation Research Progress,* Nova Science Publishers, Inc.

Van Dender, K. (2003), "Transport taxes with multiple trip purposes", *Scandinavian Journal of Economics 105*, 295-310.

Verhoef, E.T., P. Nijkamp and P. Rietveld (1996), "Second-best congestion pricing: The case of an untolled alternative*", Journal of Urban Economics 40(3),* 279-302.

Verhoef, E.T., E. Niskanen, R. Lindsey, A. de Palma, P. Moilanen, S. Proost and A. Vold (2007), "Implementation paths for marginal-cost-based pricing in urban transport: Theoretical considerations and case study results", in Jensen-Butler, C. et al. (eds.), *Road Pricing, the Economy and the Environment,* Springer Verlag, 49-78.

Verhoef, E.T. and K.A. Small (2004), "Product differentiation on roads: Constrained congestion pricing with heterogeneous users", *Journal of Transport Economics and Policy 38(1),* 127-156.

Vickrey, W.S. (1969), "Congestion theory and transport investment", *American Economic Review (Papers and Proceedings) 59*, 251-260.

Virginia Department of Transportation (2007), Tolling Facilities Report, HB 3202 (2007), December (http://www.virginiadot.org/projects/resources/TollingReportforVDOTwebsite.pdf [December 6, 2010]).

Walters, A.A. (1961), "The theory and measurement of private and social cost of highway congestion", Econometrica 29, 676-699.

Wang, J., R. Lindsey and H. Yang (2011), "Nonlinear pricing on private roads with congestion and toll collection costs", *Transportation Research Part B 45(1)*, 9-40.

Wardman, M. (2001), "A review of British evidence on time and service quality valuations", *Transportation Research E 37E,* 107-128.

Whitty, J. (2007), Oregon's mileage fee concept and road user fee pilot program: Final Report, Salem, OR: Oregon Department of Transportation, November (http://www.oregon.gov/ODOT/HWY/RUFPP/docs/RUFPP_finalreport.pdf, [September 3, 2009]).

Whitty, J. (2009), "Oregon road user fee pilot project", 2009 Symposium on Mileage-Based User Fees, April 14-15, 2009, Austin, Texas, Symposium Proceedings (http://utcm.tamu.edu/mbuf/proceedings/ [July 16, 2009]).

Wong, J-T (1997), "Basic concepts for a system for advanced booking for highway use", *Transport Policy 4(2)*, 109-114.

Yang, H. and H.-J. Huang (1998), "Principle of marginal-cost pricing: How does it work in a general road network?", *Transportation Research Part A 32(1),* 45-54.

Yang, F., and W. Y. Szeto (2006), "Day-to-day dynamic congestion pricing policies towards system optimal", Proceeding, First International Symposium on Dynamic Traffic Assignment*,* Leeds, UK, 266–275.

Yang, F., Y. Yin and J. Lu (2007), "A steepest-descent day-to-day dynamic toll", *Transportation Research Record 2039,* 83-90.

Yang, H., Q. Meng and D-H. Lee (2004), "Trial-and-error implementation of marginal-cost pricing on networks in the absence of demand functions", *Transportation Research Part B 38(6)*, 477-493.

Yin, Y. and Y. Lou (2009), "Dynamic tolling strategies for managed lanes", *Journal of Transportation Engineering 135(2),* 45-52.

Zhao, Y. and K.M. Kockelman (2006), "On-line marginal-cost pricing across networks: Incorporating heterogeneous users and stochastic equilibria", *Transportation Research Part B 40(5)*, 424-435.

Annex C: Network market conduct with atomic and non-atomic players

Refrence:

de Palma, A. and L. Engelson (2012), Network market conduct with atomic and non-atomic play, working paper.

Abstract

We consider a Stackelberg game in a static network with two routes in parallel and two user groups: a continuum of cars and a fleet of trucks. The congestion functions are affine and group specific. Each car is non-atomic and ignores the impact of his route choice on congestion. On the contrary, the coordinator of the trucks can predict the response of cars to her routing strategies. We consider several scenarios: the coordinator reducing the total travel cost of the trucks, the social optimum, the second-best optimum with coordinator attempting to minimize the total system cost, as well as the benchmark, with no coordination at all. We show that solution to the first scenario coincides with user equilibrium for small number of trucks and may improve or worsen it for large number of trucks. The set of interior solutions for the first scenario is not less than in the user equilibrium and not larger than in the second best. Finally, the route usage by trucks can be non-monotonic and perform multiple jumps when the size of the fleet increases.

1. Introduction

On the congested roads, most drivers are non-atomic individuals that are not able to cooperate in route choice. Instead of collectively designing a routing strategy that would result in a minimal total travel cost (System Optimum) they act selfishly each choosing the route with minimal average travel cost. This strategy creates negative congestion externalities that may increase the travel cost for all drivers. However for some users the driving is a part of their professional job and this creates opportunity for coordination on the part of their employer. Moreover the drivers using online navigation systems receive route suggestions that could be designed not to recommend the shortest or fastest route for the individual but such routes that would incur the lowest possible total travel cost for all navigated vehicles or for all vehicles in the transportation system. In this paper we make an attempt to answer the following questions: What is the potential of coordinating a part of the drivers to reduce the total travel cost? How large part of the congestion costs would be internalised? How this depends on the number of coordinated vehicles? How this affects the potential of the congestion pricing to reduce congestion costs?

We consider an idealised representation of transport system as a network (nodes and directed links between the nodes). Some nodes serve as origins and destinations for users making their trips through the network. The continuum of users is subdivided by a finite number of classes. The road links are endowed by positive differentiable non-decreasing cost functions that relate the cost of traversing the link to the total number of vehicles doing that. Different user classes may perceive the costs differently, i.e. may have different cost functions for the same link. The cost of a trip is a sum of the costs on all links traversed by the trip. An imaginable coordinating authority would make all users to take such routes that the total travel cost is minimised, i.e. achieve the social optimum. On the other side, the concept of user (Nash) equilibrium suggests that each user selfishly chooses a route with smallest possible cost given the route choices of all other users. It is well known that performance of congested networks under user equilibrium is suboptimal. The welfare loss due to suboptimal route choice can be measured as difference in total travel cost between the two scenarios, i. e. the benefit of coordination of route choice. Another measure of ineffectivity of user equilibrium, the price of anarchy equal to the ratio between the two cost, was introduced by Koutsoupias and Papadimitriou (1999). A large body of operations research literature is devoted to estimation of the price of anarchy. Roughgarden and Tardos (2002) considered a continuum of homogeneous users on an arbitrary congested network. They showed that 4/3 is a tight upper bound of the price of anarchy when the cost functions are linear and that the ratio has no upper bound with general differentiable increasing cost functions. Moreover, Roughgarden (2002) has shown that the highest price of anarchy attainable for a class of cost functions (possessing natural properties such as closedness w r t multiplication by a scalar) does not depend on the network topology.

One of the methods for reduction of the price of anarchy, called Stackelberg routing in the literature, suggests that a central agency, a leader, takes control of a part of the users and determines their route choice in order to minimise the total travel cost for all users. In choosing the strategy, the leader takes into account the predicted response of followers, e.g. the selfish independent users, to the strategy. We call this scenario the Second Best because it provides the minimal possible total cost when the uncontrolled users behave selfishly. Roughgarden (2004) has demonstrated that the total cost can always be reduced to precisely $1/\alpha$ times the optimal total cost by optimally controlling the share α of the total flow.

In the papers mentioned above the users are non-atomic, i.e. each user controls an infinitesimal portion of the flow and possesses no market power. Another line of research considers a finite set of atomic users, each choosing a route for travel demand for a specific origin-destination pair. Each atomic user's aim is to minimise her cost taking other users' route choices as given. In this game, each user possesses market power since she controls a substantial part of the demand. The solution sought for these games is the Nash equilibrium involving all decision makers. In the game with

unsplittable flows, each user can only use one route while the game with splittable flows allows the atomic user to distribute her demand among different routes, which is equivalent to mixed strategies. Korilis et al. (1997) consider the atomic user game on a capacitated network consisting of several parallel routes with cost functions that tend to infinity as flow approach the capacity. They showed that a leader controlling a large enough share of the demand is able to induce a system optimum. Under similar assumptions but with linear link delays, Koutsoupias and Papadimitriou (1999) derived upper and lower bounds of the price of anarchy for unsplittable atomic network game. Harks (2011) generalized the result by Korilis et al. (1997) to arbitrary semi-convex cost functions.

A series of papers considers the changes in the total cost of Nash equilibrium associated with collusions i.e. the building groups of users coordinating their route choice in order to minimise the total cost of users within the group. The atomic game with splittable flows is a special case of the game with collusions.

While all the results listed above deal with separable cost functions, Chau and Sim (2003) extend some results by Roughgarden and Tardos (2002) to symmetric cost mappings, which means that the cost of traversing a link may depend on total flows on other links, although in a symmetric manner, i.e. the Jacobian of the cost mapping is symmetric for any combination of flows. Perakis (2007) provides tight upper bounds for the price of anarchy in the case of asymmetric nonlinear cost mappings with positive semidefinite Jacobian.

Compared to these studies we go a step further in two relations. First, we include in consideration the Stackelberg strategies of network users that do not pursue to reduce the cost for the whole system but rather the cost for the group of users. Indeed, in the literature considering competition between the atomic (group) user with splittable flows and the non-atomic users behaving selfishly, the atomic users either make the first move trying to achieve the minimal travel cost for the whole system, or are assumed to take decision simultaneously with the non-atomic ones, which results in a Nash equilibrium. On the contrary, we assume in one of considered scenarios that the atomic user can predict the non-atomic users' response to their strategy and implements the strategy that delivers the best outcome for the atomic user.

Second, we consider cost mappings with indefinite Jacobian. Travel cost along a link in the network depends in our model on the total travel volume of all user groups using the link. However the cost functions are different for different user groups. As explained in subsection 3.1, this leads to an asymmetric cost mapping with indefinite Jacobian. Such cost mappings have not been exploited in the literature in relation to network games.

Brueckner (2002) considered allocation of passenger and flights between peak and off-peak periods in the context of airport congestion. He showed that the externalities arising in the free competition between the carriers are internalised when the airport is controlled by a monopolist that coordinates fares and flights in order to achieve the highest profit.

The aim of this paper is to demonstrate the effect of different routing of a part of the vehicles on the overall system cost and to study how route usage and total cost depends on the total number of coordinated vehicles. In order to avoid the problems appearing in complex network topologies, we consider a unique origin-destination pair connected by two parallel roads. We assume that there are two classes of vehicles and that the volume-cost relationships are affine with parameters differing between the routes and between the classes. The affine form of relationship between the travel demand and the average travel cost appears e.g. in the simple bottleneck model (Arnott et al., 1993). We call Second Best (SB) the Stackelberg game scenario where the leader coordinates a part of vehicles in order to reduce the total travel cost. We call Private Stackelberg(PS) the scenario when the aim of the leader is to minimize just the total travel cost of the group it controls.

We find that for small number of coordinated vehicles the PS scenario coincides with the User Equilibrium while for large number of coordinated vehicles it may reduce or increase the total travel

cost compared to the User Equilibrium, depending on the parameters of the volume-cost relationships (congestion functions). We have also found that in the User Equilibrium problem route usage by one of the two groups of vehicles may be non-monotonous in respect to the total number in this group, with no more than one jump in usage that corresponds to an interior solution of the problem. For System optimum and for Second best, we found that the behaviour of solution can be quite complex with multiple jumps between the routes. When the number of vehicles in the controlled group is large enough, the solution to the Second Best coincides with System Optimum, similar to the known case with the same congestion function for both groups.

In the next section we present our notations and the solution to the standard model with one user group. In section 3 and 4 we consider respectively User Equilibrium and System optimum for the situation with two user groups. Section 5 is devoted to the Second Best and the Stackelberg scenarios and the last section presents conclusions and suggests directions for further research.

2. The basic model and the standard results for homogenous users

In this section, we introduce notations for the simple two links network with affine congestion functions and two user groups: cars and trucks. Then, in order to prepare the base for multiclass scenarios, we consider the two well-known basic scenarios - user equilibrium and system optimum – when there are no trucks.

2.1. The basic model and notations

Consider two routes connecting the same OD-pair. The average cost functions for cars and for trucks of travelling along route r are defined respectively as $C_{v,r}\left(n_r, m_r\right) = a_r + b_r\left(n_r + m_r\right)$ and $C_{l,r}\left(n_r, m_r\right) = A_r + B_r\left(n_r + m_r\right)$ with positive parameters a_r, b_r, A_r and B_r, where n_r and m_r are the non-negative car volume and truck volume in passenger car equivalents. In order to conveniently calculate the total cost, the average costs are also defined per car equivalent, e. g. if a truck is equivalent to two passenger cars then the average cost for a truck is $2C_{l,r}$.

The total volume of cars and trucks are given respectively as

$$n_1 + n_2 = N \tag{1}$$

and

$$m_1 + m_2 = M \tag{2}$$

where $N > 0$ and $M \geq 0$.

In this paper we will consider the total cost for cars $TC_v = n_1 C_{v,1} + n_2 C_{v,2}$, the total cost for trucks $TC_l = m_1 C_{l,1} + m_2 C_{l,2}$ and the total system cost $TC_t = TC_v + TC_l$ for various scenarios: user equilibrium, system optimum, Private Stackelberg with agency controlling the trucks in order to minimise the trucks' total travel cost, and Second Best when the agency minimises the total system cost given that the car drivers behave selfishly.

Note that the cost mapping $\left(n_r, m_r\right) \to \left(C_{v,r}, C_{l,r}\right)$ has a Jacobian $\begin{pmatrix} b_r & b_r \\ B_r & B_r \end{pmatrix}$ which is indefinite since the corresponding quadratic form $b_r x_1^2 + \left(b_r + B_r\right) x_1 x_2 + B_r x_2^2 = \left(b_r x_1 + B_r x_2\right)\left(x_1 + x_2\right)$ can be positive or negative depending on the values of the variables. The price of anarchy and the price of collusion for such problem have not been considered in the literature yet.

2.2. The user equilibrium for cars

Assume $M = 0$.

In the user equilibrium, no cars use route with higher cost. If there is an interior solution,

$$0 < n_1 < N , \tag{3}$$

then the average costs are equal, i.e. $C_{v,1} = C_{v,2}$. Solving the equation and substitution into (3) implies

$$-b_2 N < a_2 - a_1 < b_1 N \tag{4}$$

which is a necessary and sufficient condition existence for interior solution. In general, the user equilibrium is unique for any combination of parameters and equal to

$$n_1^e = \begin{cases} 0 & \text{if } a_2 - a_1 \le -b_2 N, \\ \dfrac{a_2 - a_1 + b_2 N}{b_1 + b_2} & \text{if } -b_2 N < a_2 - a_1 < b_1 N, \\ N & \text{if } a_2 - a_1 \ge b_1 N. \end{cases} \tag{5}$$

Hence the total cost for all cars is

.

$$TC_v^e = \begin{cases} N\left(a_2 + b_2 N\right) & \text{if } a_2 - a_1 \le -b_2 N, \\ \dfrac{a_1 b_2 + a_2 b_1 + b_1 b_2 N}{b_1 + b_2} N & \text{if } -b_2 N < a_2 - a_1 < b_1 N, \\ N\left(a_1 + b_1 N\right) & \text{if } a_2 - a_1 \ge b_1 N. \end{cases} \tag{6}$$

2.3. The social optimum for cars

Again, assume $M = 0$.

At the social optimum, the central planner minimizes the total social cost $TC_v = n_1\left(a_1 + b_1 n_1\right) + \left(N - n_1\right)\left[a_2 + b_2\left(N - n_1\right)\right]$ subject to constraint $0 \le n_1 \le N$. For an interior solution, the marginal social costs are equalized across the routes. Solving this condition and substituting into (3) the necessary and sufficient condition for interior solution is obtained as $-2b_2 N < a_2 - a_1 < 2b_1 N$ which is weaker than (4). In general, the solution is unique for any combination of parameters and equal to

$$n_1^o = \begin{cases} 0 & \text{if } a_2 - a_1 \le -2b_2 N, \\ \dfrac{a_2 - a_1 + 2b_2 N}{2\left(b_1 + b_2\right)} & \text{if } -2b_2 N < a_2 - a_1 < 2b_1 N, \\ N & \text{if } a_2 - a_1 \ge 2b_1 N. \end{cases}$$

.

Using the definition of TC_v , the value of the minimized social cost is obtained as

$$TC_v^o = \begin{cases} N\left(a_2 + b_2 N\right) & \text{if } a_2 - a_1 \le -2b_2 N, \\ \dfrac{a_1 b_2 + a_2 b_1 + b_1 b_2 N}{b_1 + b_2} N - \dfrac{\left(a_2 - a_1\right)^2}{4\left(b_1 + b_2\right)} & \text{if } -2b_2 N < a_2 - a_1 < 2b_1 N, . \\ N\left(a_1 + b_1 N\right) & \text{if } a_2 - a_1 \ge 2b_1 N. \end{cases}$$

Comparison with equation (6) reveals the upper bound for the benefit of coordination of car routes, $BC_v \equiv TC_v^e - TC_v^o \le (a_2 - a_1)^2 / (4b_1 + 4b_2)$ which is attained for all $N \ge (a_2 - a_1)/(2b_1)$ if $a_1 \le a_2$ and for all $N \ge (a_1 - a_2)/(2b_2)$ if $a_1 > a_2$.

Figure 1 shows the car volumes on both routes in the two scenarios and the benefit of coordination as function of N for fixed values of other parameters. When N increases starting from 0, all cars first use the route with the lowest uncongested cost in both scenarios, and there is no benefit in the coordination of route choice. In the system optimum scenario, when the congestion cost reaches half of the difference between the uncongested costs, the cars gradually start using another route. In the user equilibrium scenario all cars keep using the same route, and the benefit of coordination increases. At $N = N_1$ such that the average costs with all vehicles on one route equalises with the uncongested cost at another route, the cars start using the second route in the equilibrium scenario as well. The price of anarchy is maximal at $N = N_1$. From that value of N, the usage of each route increases with the same rate in both scenarios, the difference in usage of each route between the scenarios is constant and the benefit of coordination of route choice is constant as well while the price of anarchy decreases.

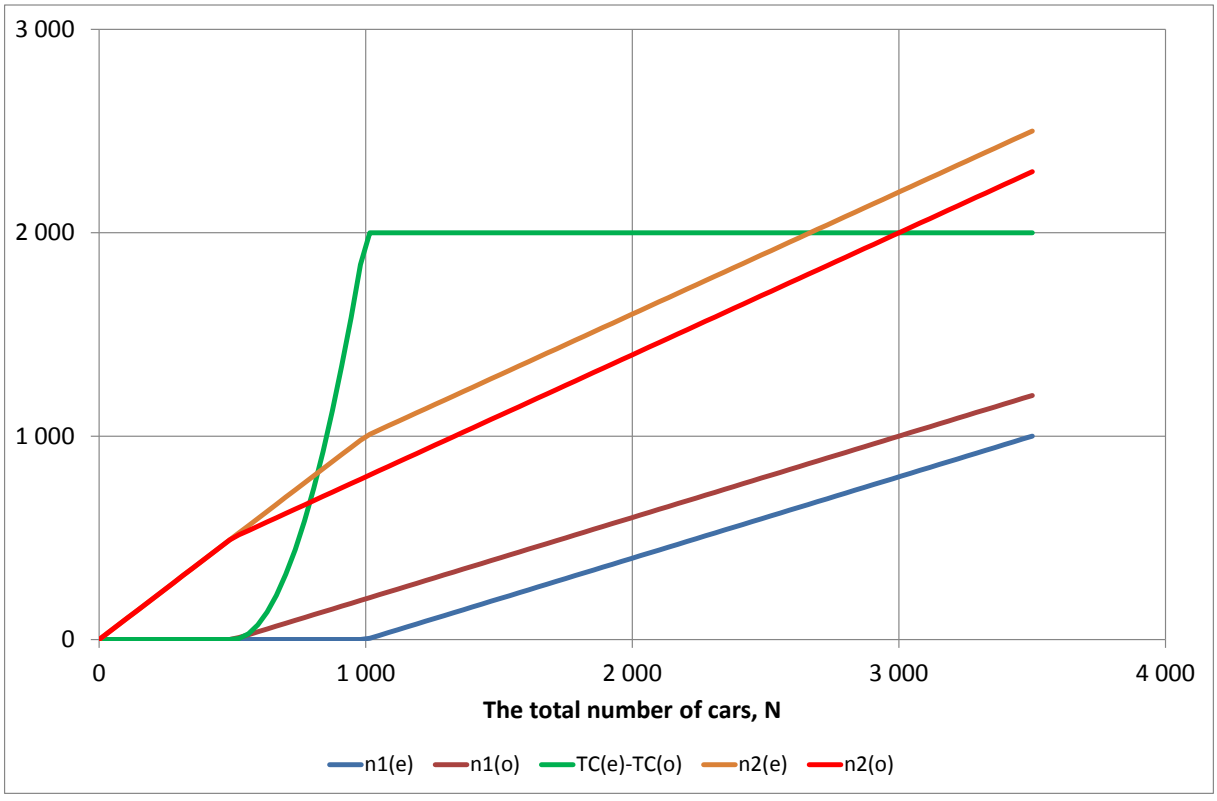


Figure 1. The route usage in the user equilibrium and the system optimum scenarios, and the benefit of coordination of cars' route choice as a function of the total number of cars. The assumed parameter values are $a_1 = 30, a_2 = 10, b_1 = 0.03, b_2 = 0.02$

3. Equilibrium with a continuum of cars and continuum of trucks

Here we consider the situation when both car volumes and truck volumes satisfy the simultaneous user equilibrium conditions, i.e. no car use route with higher cost for cars and no truck use route with higher cost for trucks. We explore the behaviour of the equilibrium when the number of trucks M

changes assuming the rest of parameters fixed and derive expressions for total travel cost for small and for large number of trucks.

3.1. Artificial optimization problem and location of the equilibria

The problem can be represented as the quadratic programming problem with the artificial objective function

$$\begin{aligned}&\min_{n_1,m_1} E\left(n_1,m_1\right)=\frac{1}{2}\left(B_1+B_2\right)\left(n_1+m_1\right)^2+\\&+\frac{B_1+B_2}{b_1+b_2}\left[a_1-a_2-b_2\left(N+M\right)\right]n_1+\left[A_1-A_2-B_2\left(N+M\right)\right]m_1\end{aligned} \tag{7}$$

subject to

$$0\le n_1\le N, 0\le m_1\le M\ . \tag{8}$$

It is easy to check that the Kuhn-Tucker conditions for the problem (7), (8) coincide with the definition of the equilibrium. Since the Jacobian of the objective function is positive semi-definite, the set of equilibria is non-empty and convex.

Given the volumes of trucks m_1 and $m_2=M-m_1$, the cars choose route according to the user equilibrium similar to section 2.2. Since the trucks incur additional fixed cost on cars equal to b_1m_1 and $b_2\left(M-m_1\right)$ along the two routes, the user equilibrium for cars is obtained by substitution $a_1+b_1m_1$ instead of a_1 and $a_2+b_2\left(M-m_1\right)$ instead of a_2 into (5). Therefore the response function of cars is

$$n_1\left(m_1\right)=\begin{cases}0, & \text{if } m_1\ge\hat{m}\\ \hat{m}-m_1, & \text{if } \hat{m}-N<m_1<\hat{m}\\ N, & \text{if } m_1\le\hat{m}-N\end{cases}\ . \tag{9}$$

where

$$\hat{m}\equiv\frac{a_2-a_1+b_2\left(N+M\right)}{b_1+b_2}. \tag{10}$$

Using the response function (9), the problem (7), (8) can be reduced to the one-dimensional problem $\min_{0\le m_1\le M} E\left(n_1^e\left(m_1\right),m_1\right)$. Substitution of (9) into (7) shows that $E\left(n_1^e\left(m_1\right),m_1\right)$ is convex quadratic function of m_1 for $m_1\le\hat{m}-N$, affine function of m_1 for $\hat{m}-N<m_1<\hat{m}$, and again convex quadratic for $m_1\ge\hat{m}$. The function $E\left(n_1^e\left(m_1\right),m_1\right)$ is illustrated in Figure 6, marked by letter E. The function is continuously differentiable (no kinks) and convex but not strictly convex. The slope of the affine part is

$$\Psi\equiv C_{l,1}\left(\hat{m}-m_1,m_1\right)-C_{l,2}\left(N-\hat{m}+m_1,M-m_1\right)=\left[A_1+B_1\hat{m}\right]-\left[A_2+B_2\left(N+M-\hat{m}\right)\right] \tag{11}$$

that is the difference of travel costs for truck on routes 1 and 2 under assumption that

$$n_1=\hat{m}-m_1\ . \tag{12}$$

Denote m' and m'' the abscissae of the vertices of the left and the right parabolas, i.e. $m'=\frac{A_2-A_1+B_2M-B_1N}{B_1+B_2}$ and $m''=\frac{A_2-A_1+B_2\left(N+M\right)}{B_1+B_2}$. It follows from the convexity of the objective that the set of solution to the problem $\min_{0\le m_1\le M} E\left(n_1^e\left(m_1\right),m_1\right)$ is convex, i.e. it is either a

singleton or an interval. The latter can only emerge if the slope $\Psi = 0$ and the interval $]\hat{m} - N, \hat{m}[$ where the function is affine overlaps with the feasible set $[0, M]$, i.e. $0 < \hat{m} < N + M$. In this case the set of solutions is the whole intersection $[\hat{m} - N, \hat{m}] \cap [0, M]$.

There are five possible cases for the location of the simultaneous user equilibrium:

Case 1. If $\Psi > 0$ then there is a unique user equilibrium with all trucks on route 2 (if $m' \leq 0$) or all cars on route 1 (if $\hat{m} - N \geq 0$).

Case 2. If $\Psi < 0$ then there is a unique user equilibrium with all trucks on route 1 (if $m'' \geq M$) or all cars on route 2 (if $\hat{m} \leq 0$).

Case 3. If $\Psi = 0$ and $\hat{m} \geq N + M$ then there is a unique common user equilibrium with all cars and trucks on route 1.

Case 4. If $\Psi = 0$ and $\hat{m} \leq 0$ then there is a unique common user equilibrium with all cars and trucks on route 2.

Case 5. If $\Psi = 0$ and $0 < \hat{m} < N + M$ then there is a continuum of user equilibria $\left(n_1^e, m_1^e\right)$ such that $\max\left(0, \hat{m} - N\right) \leq m_1^e \leq \min\left(M, \hat{m}\right)$ and $n_1^e = \hat{m} - m_1^e$. See also Proposition 3.

The five cases above can be illustrated graphically as point locations in the rectangle in Figure 2 depicting the feasible set (8). In Case 1 the solution is situated on the right edge or on the bottom edge of the rectangle, while in Case 2 it can be found on the left or on the top edge. Case 3 obtains the solution in the upper right corner and Case 4 in the lower left corner. Finally, in Case 5 there is a continuum of solutions on a line with slope -1 crossing the rectangle.

3.2. Location of the equilibrium with small number of trucks

For the purpose of comparative statics, supplement the quantities $\hat{m}$ and Ψ defined by equations (10) and (11) by argument M and note that both $\hat{m}(M)$ and $\Psi(M) = \left[A_1 + B_1 \hat{m}(M)\right] - \left[A_2 + B_2\left(N + M - \hat{m}(M)\right)\right]$ are continuous functions of M.

We will investigate the location of simultaneous user equilibrium with small number of trucks in the two cases: when cars use both routes in the situation without trucks and when just one route is used. In sections 4 and 5 we will parallel these results for the System Optimum and for the Second best scenarios and compare the total travel cost in order to see if the congestion externalities can be internalised by coordination of route choices of a part of vehicles.

Assume first that the cars use both routes when there are no trucks. Then this is true as well when the total number of trucks is small, because inequality (4) is satisfied , and (9) implies that $n_1^e(m_1) = \hat{m}(M) - m_1$ for small positive M and m_1. Substitution into the cost function for trucks gives $C_{l,1}\left(\hat{m}(M) - m_1, m_1\right) - C_{l,2}\left(N - \hat{m}(M) + m_1, M - m_1\right) = \Psi(M)$. If $\Psi(0) > 0$ then it follows from the continuity of Ψ, that, for small positive M and m_1, $C_{l,1} > C_{l,2}$ and therefore all trucks choose route 2, i. e. $m_1^e = 0$ and $n_1^e = \hat{m}(M)$. Similarly, if $\Psi(0) < 0$ then for small positive M and m_1, all trucks choose route 1, i. e. $m_1^e = M$ and $n_1^e = \hat{m}(M) - M$.

Assume now that inequality $a_2 - a_1 > b_1 N$ is fulfilled whence the cars use just route 1 when there are no trucks. Then, for small positive M and m_1,

$C_{v,2}\left(0, M-m_1\right)-C_{v,1}\left(N, m_1\right)=a_2+b_2\left(M-m_1\right)-a_1-b_1 m_1-b_1 N>0$, which implies that all cars still use route 1. Then for trucks we have $C_{l,1}=A_1+B_1\left(m_1+N\right)$ and $C_{l,2}=A_2+B_2\left(M-m_1\right)$. If $A_1+B_1 N<A_2$ then, for small M , $C_{l,1}<C_{l,2}$ and all trucks use route 1. If $A_1+B_1 N>A_2$ then, for small M , all trucks use route 2.

The results related to location of the user equilibrium with small M and the corresponding expression for total cost are summarized in Table 1. Cases when some inequality in the first column becomes equality is not analysed here because in such cases the expression for total cost with small M is dubious.

Table 1. Simultaneous user equilibrium and total cost for small number of trucks.

Conditions on parameters	Solution	Total cost
$-b_2 N<a_2-a_1<b_1 N$, $\Psi(0)>0$	$n_1^e=\hat{m}(M)$, $m_1^e=0$	$\hat{m}\left(a_1+b_1\hat{m}\right)+\left(N-\hat{m}\right)\left[a_2+b_2\left(N+M-\hat{m}\right)\right]+$ $+M\left[A_2+B_2\left(N+M-\hat{m}\right)\right]$
$-b_2 N<a_2-a_1<b_1 N$, $\Psi(0)<0$	$n_1^e=\hat{m}(M)-M$, $m_1^e=M$	$\left(\hat{m}-M\right)\left(a_1+b_1\hat{m}\right)+M\left(A_1+B_1\hat{m}\right)+$ $+\left(N+M-\hat{m}\right)\left[a_2+b_2\left(N+M-\hat{m}\right)\right]$
$a_2-a_1>b_1 N$, $A_2-A_1>B_1 N$	$n_1^e=N$, $m_1^e=M$	$N\left[a_1+b_1\left(N+M\right)\right]+M\left[A_1+B_1\left(N+M\right)\right]$
$a_2-a_1>b_1 N$, $A_2-A_1<B_1 N$	$n_1^e=N$, $m_1^e=0$	$N\left(a_1+b_1 N\right)+M\left(A_2+B_2 M\right)$
$a_2-a_1<-b_2 N$, $A_2-A_1>-B_2 N$	$n_1^e=0$, $m_1^e=M$	$N\left(a_2+b_2 N\right)+M\left(A_1+B_1 M\right)$
$a_2-a_1<-b_2 N$, $A_2-A_1<-B_2 N$	$n_1^e=0$, $m_1^e=0$	$N\left[a_2+b_2\left(N+M\right)\right]+M\left[A_2+B_2\left(N+M\right)\right]$

Results summarized in table 1 will be compared with corresponding results for other scenarios in sections 4.2 and 5.3.

3.3. Location of the equilibrium with large number of trucks

Let's now investigate what will be the solution and the cost when the number of trucks is large. First, substitute (10) into (11) to obtain

$$\Psi\left(M\right)=\frac{\Delta B_1 B_2\left(N+M\right)}{b_1+b_2}+\frac{\left(B_1+B_2\right)\left(a_2-a_1\right)}{b_1+b_2}-\left(A_2-A_1\right), \tag{13}$$

where

$$\Delta \equiv \frac{b_2}{B_2}-\frac{b_1}{B_1} \tag{14}$$

is the difference in relative externalities.

Equation (13) implies that $\Psi\left(M\right)$ is an affine function of M and, if $\Delta \neq 0$, has the same sign as Δ for large values of M . Moreover, as the following proposition shows, the solution for M large

enough can only be located on the left or on the right edge of the rectangle in Figure 2, i. e. all cars are on the same route. This means that for large M the trucks crowd all cars out to one route and equalise their cost between the two routes.

Proposition 1. Assume $\Delta \neq 0$ and $M > \max\left(\frac{A_1 - A_2 + B_1 N}{B_2}, \frac{A_2 - A_1 + B_2 N}{B_1}, \frac{(A_2 - A_1)(b_1 + b_2) - (a_2 - a_1)(B_1 + B_2)}{\Delta B_1 B_2} - N \right)$. Then the equilibrium is unique and involves partial separation with all cars on one route and trucks on both routes. Namely, if $\Delta > 0$, then $\left(n_1^e, m_1^e\right) = \left(N, \frac{A_2 - A_1 + B_2 M - B_1 N}{B_1 + B_2} \right)$, and the total cost

$$TC_t^e = N\left(a_1 + b_1 N\right) + M\left(A_2 + B_2 M\right) + \frac{A_2 - A_1 + B_2 M - B_1 N}{B_1 + B_2}\left(b_1 N - B_2 M\right), \tag{15}$$

and if $\Delta < 0$, *then* $\left(n_1^e, m_1^e\right) = \left(0, \frac{A_2 - A_1 + B_2 (N + M)}{B_1 + B_2} \right)$, *and the total cost.*

$$TC_t^e = N\left(a_2 + b_2 N\right) + M\left(A_1 + B_1 M\right) + \frac{A_1 - A_2 + B_1 M - B_2 N}{B_1 + B_2}\left(b_2 N - B_1 M\right). \tag{16}$$

Proof. See Appendix A.

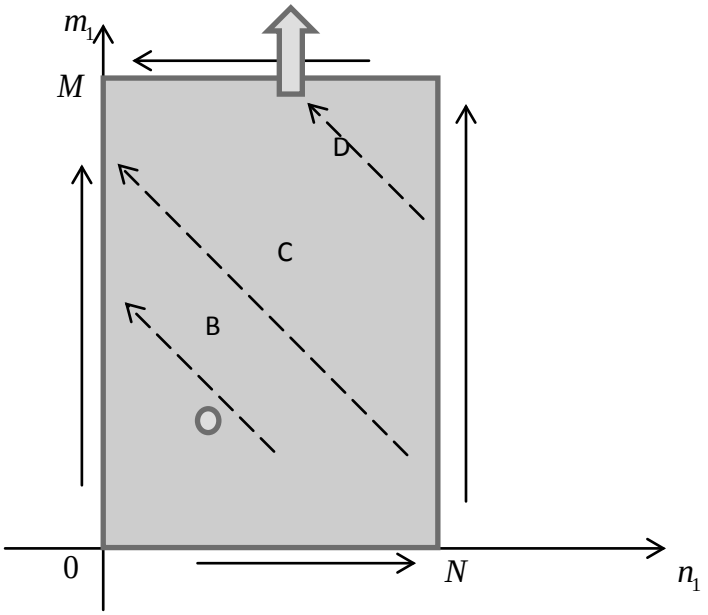


Figure 2. The solution domain for user equilibria in the two mode two route network. The small circle represents the equilibrium at $M = 0$. As M increases the rectangle becomes taller. Possible evolution of the equilibria is shown by the arrows. The possible kinds of jumps B, C, and D (for $\Delta < 0$) are shown by the dashed arrows.

3.4. Evolution of the equilibrium with increasing number of trucks

Now consider what happens with the user equilibrium for intermediate number of trucks (not very small and not very large) . Note that the solution $\left(n_1^e, m_1^e\right)$ is continuous in respect to M as long as $\Psi(M) \neq 0$. This can be shown e.g. by writing down all expressions for the solution in the cases 1 to 5 of subsection 3.1.

Assume $\Delta \neq 0$. Since interior solution can only exist for one value of $M = M_1$ satisfying equation $\Psi(M_1) = 0$, the equilibrium continuously moves along the edges of the rectangle for $M \neq M_1$. Note that the rectangle becomes higher as M increases. The evolution of solution as M increases can still be very diverse. In order to reduce the number of cases to consider, let's make the following two assumptions.

Assumption 1. When there are no trucks, the cars use both routes in the user equilibrium, i.e. condition (4) is fulfilled: $-b_2 N < a_2 - a_1 < b_1 N$.

This assumption implies that neither $(0,0)$ nor (N,M) can be a solution, i.e. the cases 3 and 4 of subsection 3.1 cannot occur for any M. Moreover, as long as M is so small that the trucks cannot crowd out all cars to the same route, i.e. $M < \min\left[\frac{a_2 - a_1 + b_2 N}{b_1}, \frac{a_1 - a_2 + b_1 N}{b_2}\right]$, the route taken by the trucks in the equilibrium is determined by the sign of $\Psi(M)$: if $\Psi(M)$ is positive (negative) then all trucks are on route 2 (route 1).

Graphically, the vertical axe crosses the affine part of function E is Figure 6 when $M = 0$.

Assumption 2. When there are no trucks and the cars choose routes according to the user equilibrium, route 2 is cheaper for trucks than route 1.

Algebraically, Assumption 2 can be expressed as $C_{l,2}(n_1^e, 0) < C_{l,1}(n_2^e, 0)$ or as $\Psi(0) > 0$. Graphically, the slope of the affine part of function E is Figure 6 is positive when $M = 0$.

Assumptions 1 and 2 together imply that for small total number of trucks M all trucks take route 2, i.e. the user equilibrium is represented by a point on the bottom edge (and not in a corner) of the rectangle as shown in Figure 2, $(n_1^e, m_1^e) = \left(\frac{a_2 - a_1 + b_2(N+M)}{b_1 + b_2}, 0\right)$. This corresponds to the first row of Table 1. As M increases, the new trucks accumulate on route 2 and increase the cost for cars on this route, whence the cars move to route 1, i.e. the solution point moves to the right along the bottom edge of the rectangle. Indeed, $\frac{\partial n_1^e}{\partial M} = \frac{b_2}{b_1 + b_2} > 0$.

The value Δ is significant because it determines evolution of the solution for higher values of M as shown in Proposition 1.

Consider first the case $\Delta > 0$. Then $\Psi(M) > 0$ for all values of M i.e. the slope of the affine part of function E in Figure 6 is positive and increases. As M increases starting from 0, the solution moves towards the right bottom corner $(N,0)$ where it may stay for an interval of values of M. At this point the separation takes place with all trucks on route 2 and all cars on route 1. As M increases further, the solution leaves the corner and moves up along the right edge, which means that the number of trucks on route 1 increases, while the cars still all use route 1. Note that the number of truck on route 2 increases as well, so m_1 increases slower than M.

If $\Delta < 0$ then $\Psi(M)$ is positive for $M < M_1$ and negative for $M > M_1$. Graphically, the slope of the affine part of function E in Figure 6 decreases. Again, as M increases from 0, the solution moves right along the bottom edge of the rectangle in Figure 2 and may continue through the separation

corner $(N,0)$ up along the right edge. At $M = M_1$, where $\Psi(M_1) = 0$, the solution jumps to another edge of the rectangle. Depending on the values of other parameters, this may happen before or after the solution reaches the corner. Actually, four different types of jumps may occur: (A) from the bottom edge to the top edge, (B) from the bottom edge to the left edge, (C) from the right edge to the left edge, or (D) from the right edge to the top edge of the rectangle. According to equation (12), the jump does not change the total volume of cars and trucks on route 1. The jump types (B), (C), and (D) are depicted by the dashed arrows in Figure 2. The jump of type (A) could not be depicted there because it can only occur with $N > M$.

If the solution jumps to the top edge then all trucks switch to route 1. As M increases further, all new trucks take route 1 and gradually crowd the cars out to route 2, hence the solution moves left along the top edge (which gradually shifts up). In the left upper corner $(0,M)$, the separation occurs with all trucks on route 1 and all cars on route 2. As M increases further, the solution falls behind the separation corner and remains on the left edge for larger values of M. If the jump occurs to the left edge then the solution still remains there for larger values of M. With increasing M, the solution remains on the left edge and moves up, slower than M.

Figure 3 shows how the equilibrium car usage n_1^e and the equilibrium truck usage m_1^e of route 1 changes with M for different types of the jumps.

Note that the truck usage of a route may be non-monotonous in respect to the total number M of trucks. For example, the jump of type (C) means that m_2^e increases with rate $\frac{B_1}{B_1 + B_2}$ before the jump, drops by N at $M = M_1$, and continues to increase with rate $\frac{B_1}{B_1 + B_2}$ after the jump. In the interior equilibrium, at $M = M_1$, both the trucks and the car equalise their costs across the routes. For M just below and just above M_1 the trucks' costs on both routes are the same across the routes and almost the same as at M_1. If $\Delta < 0$, then for M just below M_1 the car cost on route 1 is lower than that on route 2, while for M just above M_1 the car cost on route 1 is higher than that on route 2. Therefore the cars move all N vehicles from route 1 to route 2 when M passes M_1. In order to keep the same route costs, the trucks compensate by moving N vehicles from route 2 to route 1.

The non-monotonicity of route usage may have important implications for transportation planning because it demonstrates that simple growth factor methods sometimes used for forecast of traffic flows by different vehicle types may come out very wrong.

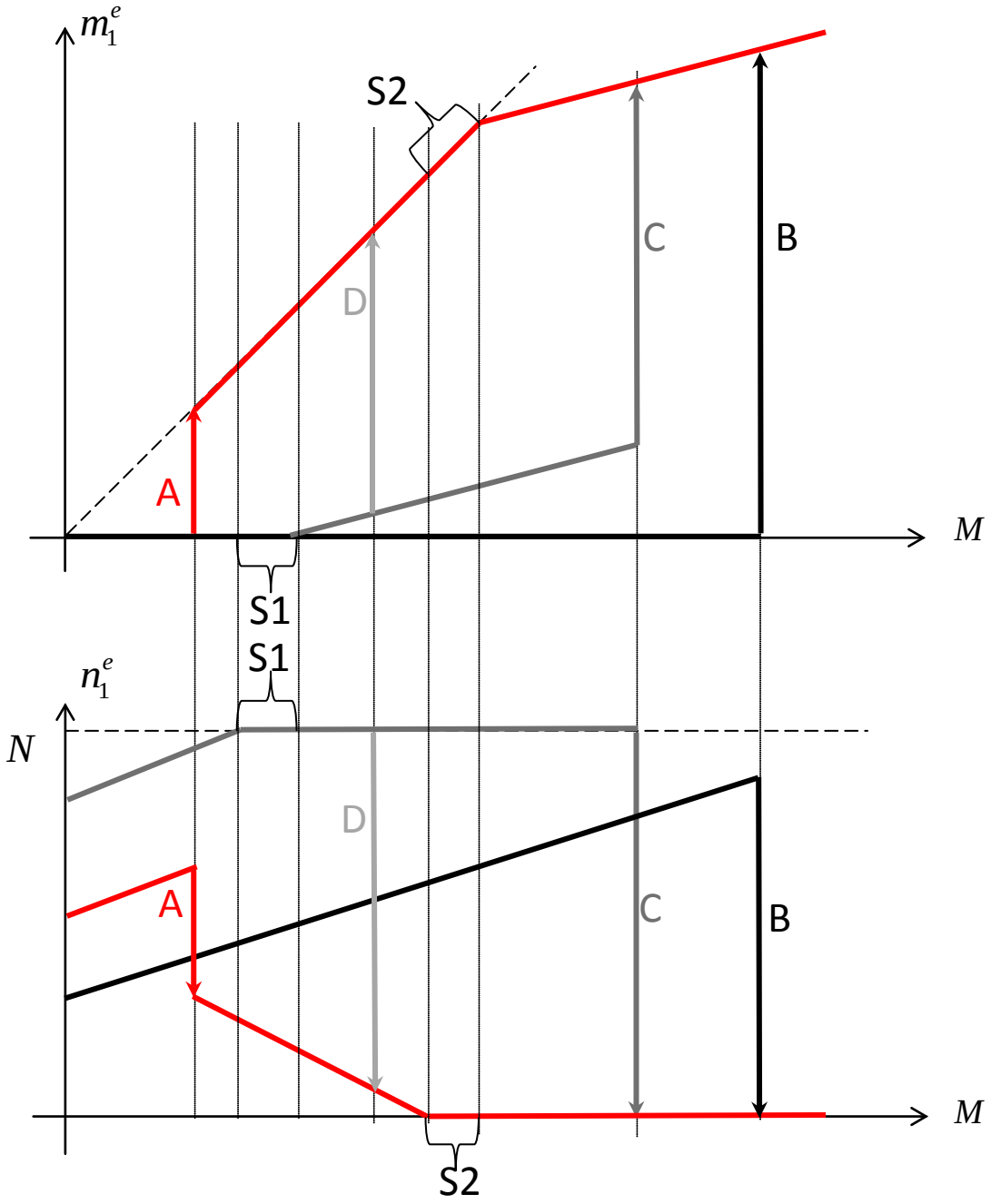


Figure 3. Evolution of the equilibrium solution with increasing total number of trucks in four cases satisfying Assumptions 1 and 2 and Δ<0. The upper (lower) plot shows the usage of route 1 by trucks (cars). Direction of the four kinds of jumps A, B, C and D is shown by the arrows. The separation S1 takes place when all cars are on route 1 and all trucks on route 2. It ensues with cases C, and D. The separation S2 occurs with all cars on route 2 and all trucks on route 1. It can be combined with cases A and D.

4. System optimum for cars and trucks

In this section we derive expressions for the total travel cost when the number of trucks M is small enough or large enough and compare with the expressions obtained for user optimum in the last section. We also demonstrate evolution of the system optimum solution with increasing M for a numerical example and compare it with the user equilibrium.

4.1. Cardinality and location of the system optimum

The system optimum suggests that both cars' and trucks' route choice are coordinated so that the total cost

$$\begin{aligned} TC_t\left(n_1,m_1\right) &= n_1\left[a_1+b_1\left(n_1+m_1\right)\right]+\left(N-n_1\right)\left[a_2+b_2\left(N+M-n_1-m_1\right)\right]+ \\ &+m_1\left[A_1+B_1\left(n_1+m_1\right)\right]+\left(M-m_1\right)\left[A_2+B_2\left(N+M-n_1-m_1\right)\right] \end{aligned} \tag{17}$$

is minimized subject to the feasibility constraints (8). Given the volumes of trucks m_1 and $m_2 = M - m_1$, the response of cars is obtained by minimization of $TC_t\left(n_1,m_1\right)$ subject to $0 \le n_1 \le N$ which gives

$$n_1^o\left(m_1,M\right) = \begin{cases} 0, & \text{if } m_1 \ge \breve{m}\left(M\right) \\ \delta\left(\breve{m}\left(M\right)-m_1\right), & \text{if } \breve{m}\left(M\right)-N/\delta < m_1 < \breve{m}\left(M\right) \\ N, & \text{if } m_1 \le \hat{m}\left(M\right)-N/\delta \end{cases} \tag{18}$$

where

$$\breve{m}\left(M\right) = \frac{b_2\left(2N+M\right)+B_2M-a_1+a_2}{b_1+b_2+B_1+B_2} \tag{19}$$

and

$$\delta = \frac{b_1+b_2+B_1+B_2}{2\left(b_1+b_2\right)} \tag{20}$$

Substitution into (17) obtains the one-dimensional minimization problem $\min_{0\le m_1\le M} \Gamma\left(m_1,M\right) = TC_t\left(n_1^o\left(m_1,M\right),m_1\right)$, where, for any M , the objective is a convex quadratic function of m_1 for $m_1 \le \breve{m}\left(M\right)-N/\delta$, concave quadratic for $\breve{m}\left(M\right)-N/\delta < m_1 < \breve{m}\left(M\right)$ and again convex quadratic for $m_1 \ge \breve{m}\left(M\right)$. The plot of the function is shown on Figure 6, marked by Γ. Equation (18) implies that interior solutions can only exist with m_1 in the second interval $\left]\breve{m}\left(M\right)-N/\delta,\breve{m}\left(M\right)\right[$. On this interval, the second derivative $\frac{\partial\Gamma\left(m_1,M\right)}{\partial m_1} = -\frac{\left(B_1+B_2-b_1-b_2\right)^2}{4\left(b_1+b_2\right)}$. Therefore, interior solution cannot exist if

$$b_1+b_2 \ne B_1+B_2\,. \tag{21}$$

Ìn the sequel we will assume that inequality (21) is satisfied.

Thus, due to the properties of function Γ, there is normally just one solution $\left(n_1^o,m_1^o\right)$ to the system optimum problem but for some values of M two solutions may appear. Related to Figure 2, there is either a solution on an edge of the rectangle or two solutions on different edges.

Location of the system optimum with small *M*Since the system optimum can be considered as user equilibrium with average costs replaced by marginal costs, the cars choose route with the lowest marginal cost and so do the trucks. The marginal cost for trucks on route r, $MC_{l,r} = \left(A_r+B_rn_r+B_rm_r\right)+B_rm_r+b_rn_r$, includes the average cost for a truck (in the parenthesis) and the congestion cost that the truck incurs on other trucks (the second term) and on the cars (the third term) using the same route.

Assume that the cars use both routes in the system optimum when there are no trucks. Then $-2b_2N < a_2 - a_1 < 2b_1N$ and, similar to the user equilibrium scenario, the cars use both routes for small M and m_1 as well. The difference between marginal route costs for trucks,

$$\begin{aligned}&\Phi(M, m_1) = MC_{l,1} - MC_{l,2} = \\ &= A_1 + (B_1 + b_1) n_1^o(m_1, M) + 2B_1 m_1 - A_2 - (B_2 + b_2)\left[N - n_1^o(m_1, M)\right] - 2B_2(M - m_1)\end{aligned} \tag{22}$$

has the same sign for small for small enough M and m_1 as for $m_1 = M = 0$. If it is positive, i.e. $\Phi(0,0) = \frac{\Delta B_1 B_2 N}{b_1 + b_2} + \frac{(b_1 + b_2 + B_1 + B_2)(a_2 - a_1)}{2(b_1 + b_2)} + A_1 - A_2 > 0$ then the solution is $m_1^o = 0$, $n_1^o = \delta\breve{m}(M)$. Alternatively, if $\Phi(0,0) < 0$ then $m_1^o = M$ and $n_1^o = \delta\left[\breve{m}(M) - M\right]$.

Now, assume instead the inequality $a_2 - a_1 > 2b_1N$. Then all cars are on route 1 for small M. Then all trucks use route 1 if $MC_{l,1} - MC_{l,2} = A_1 + (B_1 + b_1)N - A_2 < 0$ and route 2 if $A_1 + (B_1 + b_1)N - A_2 > 0$. Table 2 presents all cases and the total cost expressions.

Table 2. Simultaneous system optimum and total cost expressions for small number of trucks.

Conditions on parameters	Solution	Total cost
$-2b_2N < a_2 - a_1 < 2b_1N$, $\Phi(0,0) > 0$	$n_1^o = \delta\breve{m}(M)$, $m_1^o = 0$	$\delta\breve{m}(a_1 + b_1\delta\breve{m}) + M\left[A_2 + B_2(N + M - \delta\breve{m})\right] + (N - \delta\breve{m})\left[a_2 + b_2(N + M - \delta\breve{m})\right]$
$-2b_2N < a_2 - a_1 < 2b_1N$, $\Phi(0,0) < 0$	$n_1^o = \delta\left[\breve{m}(M) - M\right]$, $m_1^o = M$	$\delta(\breve{m} - M)\left[a_1 + b_1(\delta\breve{m} - \delta M + M)\right] + (N + \delta M - \delta\breve{m})\left[a_2 + b_2(N + \delta M - \delta\breve{m})\right] + M\left[A_1 + B_1(\delta\breve{m} - \delta M + M)\right]$
$a_2 - a_1 > 2b_1N$, $A_2 - A_1 > (B_1 + b_1)N$	$n_1^o = N$, $m_1^o = M$	$N\left[a_1 + b_1(N + M)\right] + M\left[A_1 + B_1(N + M)\right]$
$a_2 - a_1 > 2b_1N$, $A_2 - A_1 < (B_1 + b_1)N$	$n_1^o = N$, $m_1^o = 0$	$N(a_1 + b_1N) + M(A_2 + B_2M)$
$a_2 - a_1 < -2b_2N$, $A_2 - A_1 > -(B_2 + b_2)N$	$n_1^o = 0$, $m_1^o = M$	$N(a_2 + b_2N) + M(A_1 + B_1M)$
$a_2 - a_1 < -2b_2N$, $A_2 - A_1 < -(B_2 + b_2)N$	$n_1^o = 0$, $m_1^o = 0$	$N\left[a_2 + b_2(N + M)\right] + M\left[A_2 + B_2(N + M)\right]$

Comparison between Table 1 and Table 2 is cumbersome because set of parameter values shown in the first column of Table 1 in many cases overlap with but are not contained in the sets in the first column of Table 2 or vice versa. However, the third (excluding the headers) and the last parameter set in Table 2 is contained in the corresponding set of Table 1, and the expression for the total cost is

the same. This means that if the marginal cost of one route is dominated by marginal cost of another route for both trucks and cars when there are no trucks and all cars drive on the first route then for small M the user equilibrium coincides with the system optimum. Comparison of the forth and the fifth rows shows that the same is true if the marginal cost of one route is less than the marginal cost of another route for cars but the average cost of the first route for trucks is higher than for the second route when there are no trucks and all cars drive on the first route. If in the user equilibrium without trucks the cars use both routes and both the average and the marginal cost for trucks is higher on route 1 than on route 2 then the benefit of coordination of route choice for small M can be calculated based on the cost expressions in the first rows of the tables, namely

$$TC_t^e - TC_t^o = \frac{\left[a_2 - a_1 + (b_2 - B_2)M\right]^2}{4(b_1 + b_2)}. \tag{23}$$

4.2. Location of the system optimum and the benefit of coordination of truck routes with large number of trucks

Similar to the user equilibrium, the quantity Δ defined in (14) determines where the solution is located for large M .

Proposition 2. *If* $\Delta > 0$ *then for*

$$M > \max\left[\frac{A_1 - A_2 + (b_1 + B_1)N}{2B_2}, \frac{2b_2N - a_1 + a_2}{b_1 + B_1}, \frac{(b_1 + b_2 + B_1 + B_2)\left[A_2 - A_1 + (b_2 + B_2)N\right] + 2(B_1 + B_2)(2b_1N + a_1 - a_2)}{2\Delta B_1 B_2}\right]$$

the social optimum is unique and involves partial separation with all cars on route 1 and trucks on both routes, $(n_1^o, m_1^o) = \left(N, \frac{A_2 - A_1 - (b_1 + B_1)N + 2B_2M}{2(B_1 + B_2)}\right)$*, and the total cost is*

$$TC_t^o = N(a_1 + b_1N) + M(A_2 + B_2M) - \frac{\left[A_2 - A_1 - (b_1 + B_1)N + 2B_2M\right]^2}{4(B_1 + B_2)}. \tag{24}$$

If $\Delta < 0$ *then for*

$$M > \max\left[\frac{A_2 - A_1 + (b_2 + B_2)N}{2B_1}, \frac{2b_1N - a_2 + a_1}{b_2 + B_2}, \frac{(b_1 + b_2 + B_1 + B_2)\left[A_2 - A_1 - (b_1 + B_1)N\right] - 2(B_1 + B_2)(2b_2N - a_1 + a_2)}{2\Delta B_1 B_2}\right]$$

the social optimum is unique and involves partial separation with all cars on route 2 and trucks on both routes, $(n_1^o, m_1^o) = \left(0, \frac{A_2 - A_1 + (b_2 + B_2)N + 2B_2M}{2(B_1 + B_2)}\right)$*, and the total cost*

$$TC_t^o = N(a_2 + b_2N) + M(A_1 + B_2M) - \frac{\left[A_1 - A_2 + (b_2 + B_2)N + 2B_1M\right]^2}{4(B_1 + B_2)}. \tag{25}$$

Proof. See Appendix A. Comparison of (24) with (15) and (25) with (16) obtains the expressions for the benefit of complete coordination of route choices for M large enough, namely

$$TC_t^e - TC_t^o = \frac{\left[A_2 - A_1 + (b_1 - B_1)N\right]^2}{4(B_1 + B_2)} \text{ if } \Delta > 0 \tag{26}$$

and

$$TC_t^e - TC_t^o = \frac{\left[A_2 - A_1 + (B_2 - b_2)N\right]^2}{4(B_1 + B_2)} \text{ if } \Delta < 0. \tag{27}$$

The last expression could be obtained from (23) by interchanging the roles of cars and trucks. This is natural because the scenario is symmetric in respect to cars and trucks and both equations are related to the situation when one of the groups is much larger than the other.

Note that the benefit of coordination becomes constant for large values of M. Moreover, the coordination gives no benefit if the uncongested costs for trucks are the same on both routes and, for each route, the congestion parameters are the same for cars and trucks.

4.3. Evolution of the system optimum with increasing number of trucks

Similar to the user equilibrium, we consider now the evolution of solution with increasing M and fixed N. Again, the truck usage of route 1 as a function of the total number of trucks is piecewise affine. Assumption 1, i.e. condition (4), is sufficient (but not necessary) to guarantee that the cars use both routes in the system optimum when there are no trucks. However Assumption 2 needs to be slightly modified. We therefore introduce

Assumption 2'. When there are no trucks and the cars choose routes according to the system optimum, route 2 has lower marginal cost for trucks than route 1.

Assumption 2s can be expressed as $\Phi(0,0) > 0$ where Φ is defined by equation (22).

The evolution of the system optimum solution when M increases from 0 under Assumptions 1 and 2' can be similar to the user equilibrium solution under Assumptions 1 and 2, see subsection 2.3.2. However there can be more jumps, depending on the values of the parameters. An important difference compared to the user equilibrium is that the interior solution does not need to exist when the jump to another edge of the rectangle occurs. The jump occurs when the minimal total travel cost over the bottom and the right edges of the rectangle in Figure 2 equalizes with the minimal total travel cost over the left and the top edges, i.e. when there are two solutions to the system optimum problem. Of course, the values of M for which the solution transfers between the edges may not coincide with the user equilibrium scenario. A case when there are two jumps of the system optimum solution is illustrated in Figure 4. In the user equilibrium, all trucks use route 1 for $M \leq 4$, then the trucks are divided between the two routes, and then all truck switch to route 2 at $M = 14$. The System optimum suggests that all trucks use route 2 for small M, switch to route 1 at $M = 1.7$ and back to route 2 at $M = 4.2$. The truck usage of route 1 starts to increase again at $M = 84.7$ in the system optimum scenario and at $M = 147$ in the user equilibrium scenario. This example shows that the system optimum trajectory can be rather different from the user equilibrium and may have multiple jumps.

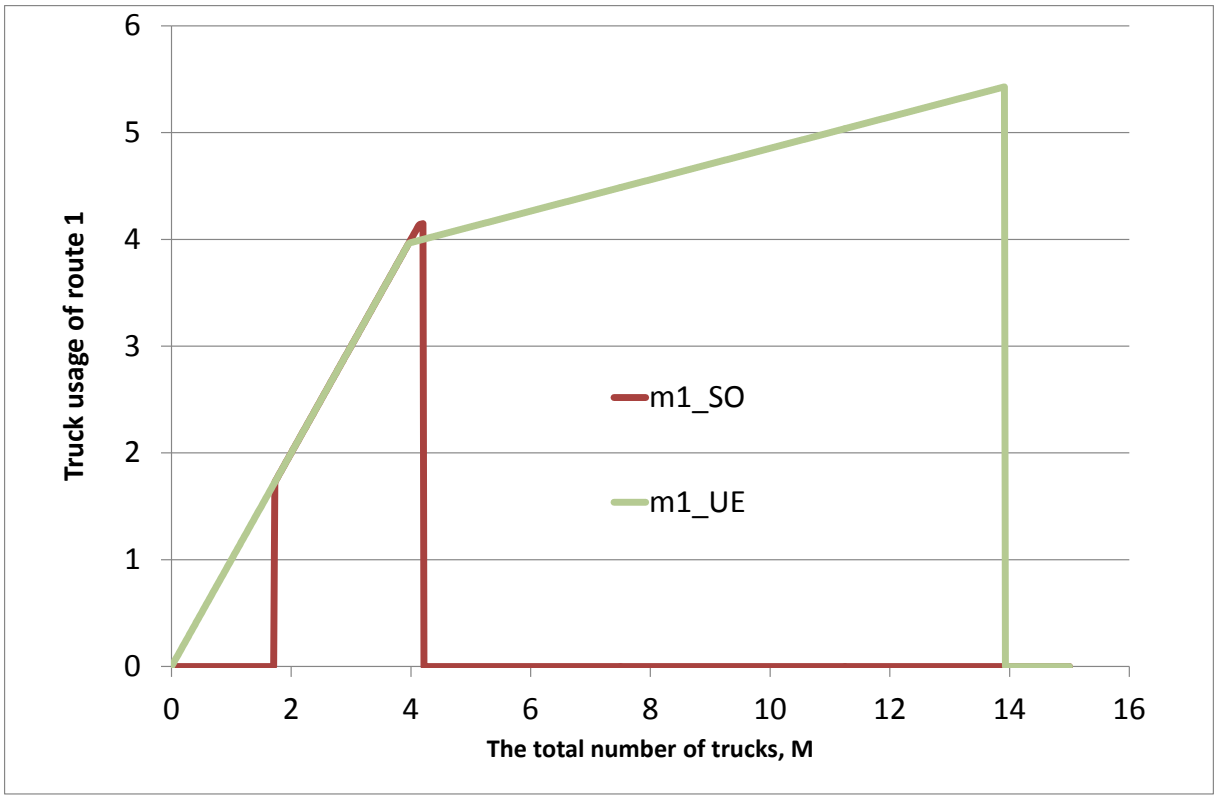


Figure 4. Example with two consecutive jumps of truck usage of route 1 in the system optimum scenario (m1_SO) and the corresponding evolution of the user equilibrium (m1_UE) as the total number of trucks increases. The assumed parameter values are

$a_1 = 30, a_2 = 1.4, b_1 = 0.9, b_2 = 1, A_1 = 3, A_2 = 1, B_1 = 5.8, B_2 = 1, N = 25$

Evolution of both user equilibrium and system optimum solutions and the benefit of coordination of route choices as functions of M are illustrated in Figure 5 for the same values of parameters. For small M, all trucks take route 2 in both scenarios, and the selfish cars overuse route 1 compared to the system optimum. As number of trucks on route 2 increases, the cars move to route 1 in both scenarios and eventually the user equilibrium coincides with system optimum. As number of trucks on route 2 increases further, the cars underuse route 1 in the user equilibrium. Just before the jump of the user equilibrium solution the highest benefit of coordination of route choices is obtained over all values of M. The jump of user equilibrium solution does not change the benefit of coordination since it does not change the total flow and the costs on any route. For the chosen values of parameters, the jump is of kind B, from the bottom to the left edge of the rectangle in Figure 2. The jump of social optimum occurs for higher value of M than the jump of user equilibrium. The minimal cost over the bottom and the right edges of rectangle in Figure 2 coincides here with the minimal cost over the top and the left edge, therefore the benefit of coordination of route choice is continuous in respect to M. After the second jump, all cars in both scenarios use route 2 while the trucks use both routes. The number of trucks increases on both routes when M increases further, and the rate of increase is the same for user equilibrium and for social optimum. The benefit of coordination is due to the coordination of trucks and is constant after the latter jump.

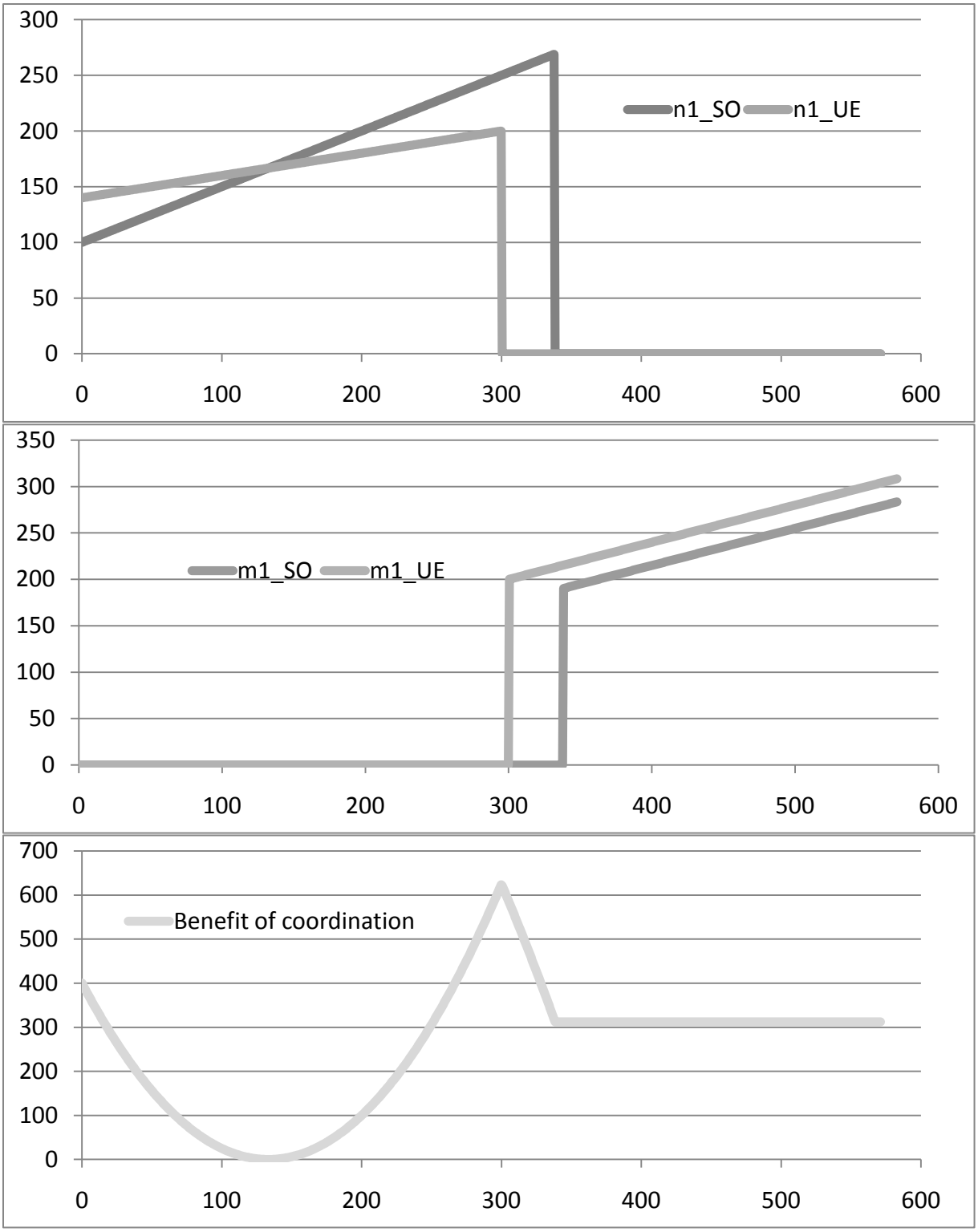


Figure 5. Usage of route 1 by cars (a) and by trucks (b) in the User Equilibrium and in the System Optimum scenarios, and the benefit of route choice coordination (c) as functions of the total number of trucks. The assumed parameter values are

$a_1 = 10,\ a_2 = 30,\ b_1 = 0.2,\ b_2 = 0.05,\ A_1 = 120,\ A_2 = 100,\ B_1 = 0.3,\ B_2 = 0.2,\ N = 300$

5. Equilibrium when Trucks are coordinated

This means that the trucks are run by an agency which sets the route proportions for the trucks in order to achieve the minimal total cost for trucks (private agency) or the minimal total cost for trucks

and cars (governmental agency). We will call the first scenario “Private Stackelberg” (PS) and “Social Second Best” (SB) correspondingly. The SB is known in the literature as “Stackelberg routing”. The agency is a leader, i.e. it takes into account the route choice behaviour of cars and how it will be affected by the agency’s decision. Each car driver is selfish price-taker and chooses the route taking into account the own travel costs only. Lacking coordination, the car drivers do not consider themselves able to influence the routes of trucks or other cars.

Compared to the User Equilibrium, the behaviour of cars is the same, that is they choose the route with lower average cost. However trucks in the SB scenario always choose the route with lower marginal cost while in the PS scenario they choose the route with lower partial marginal cost. This partial marginal cost includes the congestion cost the trucks incur on other trucks but not the congestion cost they incur on cars.

5.1. The optimization problem

Given flows of trucks $m_1, m_2 = M - m_1$, the travel costs for cars are $C_{v,1} = a_1 + b_1\left(n_1 + m_1\right)$ and $C_{v,2} = a_2 + b_2\left(N - n_1 + M - m_1\right)$. Solving for the user equilibrium for cars, one obtains the response function similar to (9),

$$n_1\left(m_1\right) = \begin{cases} 0, & \text{if } m_1 \geq \hat{m} \\ \hat{m} - m_1, & \text{if } \hat{m} - N < m_1 < \hat{m} \\ N, & \text{if } m_1 \leq \hat{m} - N \end{cases} \tag{28}$$

where $\hat{m}$ is defined by (10).

The total cost of trucks and the total cost of cars are

$$TC_l\left(n_1\left(m_1\right), m_1\right) = m_1\left[A_1 + B_1\left(n_1\left(m_1\right) + m_1\right)\right] + \left[M - m_1\right]\left[A_2 + B_2\left(N + M - n_1\left(m_1\right) - m_1\right)\right]\}$$

and

$$TC_v\left(n_1\left(m_1\right), m_1\right) = n_1\left(m_1\right)\left[a_1 + b_1\left(n_1\left(m_1\right) + m_1\right)\right] + \left[N - n_1\left(m_1\right)\right]\left[a_2 + b_2\left(N + M - n_1\left(m_1\right) - m_1\right)\right]$$

respectively. In the PS scenario, m_1 is chosen so that the total cost for trucks is minimized, $\min_{0 \leq m_1 \leq M} TC_l\left(n_1\left(m_1\right), m_1\right)$. In the SB scenario, the trucks are controlled so as to minimize the total cost for all vehicles, i.e. $\min_{0 \leq m_1 \leq M} TC_t\left(n_1\left(m_1\right), m_1\right) = TC_l\left(n_1\left(m_1\right), m_1\right) + TC_c\left(n_1\left(m_1\right), m_1\right)$. Substitution of (28) into the expression for TC_l above obtains that $TC_l\left(n_1\left(m_1\right), m_1\right)$ is a convex quadratic function of m_1 for $m_1 \leq \hat{m} - N$, affine function of m_1 for $\hat{m} - N < m_1 < \hat{m}$, and again convex quadratic for $m_1 \geq \hat{m}$. The same is true for $TC_t\left(n_1\left(m_1\right), m_1\right)$ since $TC_c\left(n_1\left(m_1\right), m_1\right)$ is affine increasing for $m_1 \leq \hat{m} - N$, constant for $\hat{m} - N < m_1 < \hat{m}$, and affine decreasing for $m_1 \geq \hat{m}$.

The slope of the affine part of both $TC_l\left(n_1\left(m_1\right), m_1\right)$ and $TC_t\left(n_1\left(m_1\right), m_1\right)$ is Ψ as defined by equation (13). Note that the objective functions $TC_l\left(n_1\left(m_1\right), m_1\right)$ and $TC_t\left(n_1\left(m_1\right), m_1\right)$ can have downward or upward kinks at $m_1 = \hat{m} - N$ and at $m_1 = \hat{m}$. The slope of the left and the right affine parts of $TC_c\left(n_1\left(m_1\right), m_1\right)$ are $b_1 N$ and $-b_2 N$ respectively. For the richest case when $0 < \hat{m} - N$ and $\hat{m} < M$, all three total cost functions are illustrated in Figure 6.

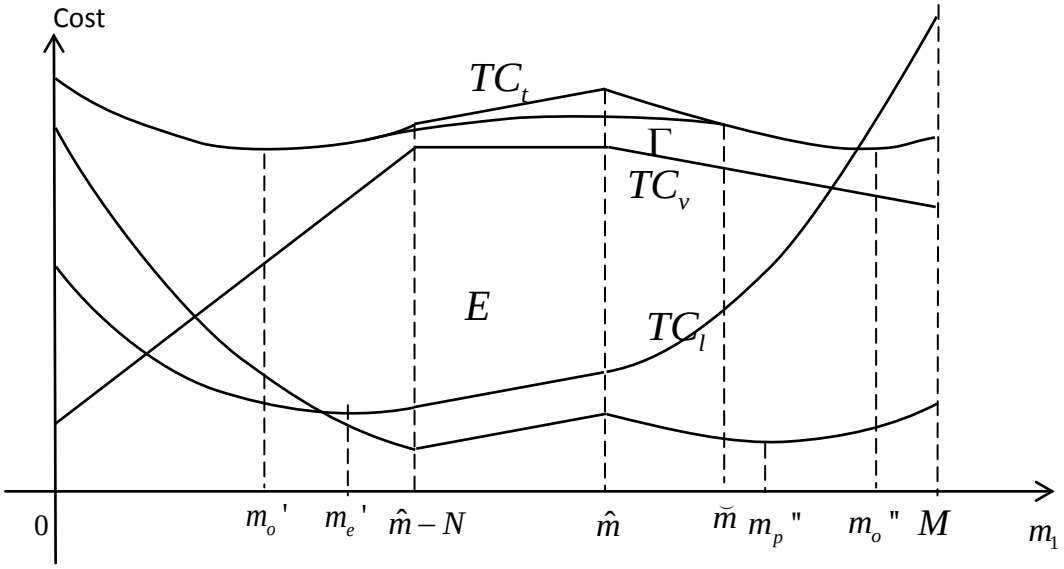


Figure 6. The total costs as functions of the truck usage of route 1 and the candidate minimum cost solutions for route decisions of trucks. The curves TC_l, TC_v, and TC_t plot respectively the total cost for trucks, the total cost for cars, and the total cost for all vehicles, under condition that the cars choose the route selfishly without coordination; in particular, the cars all choose route 1 if $m_1 < \hat{m} - N$ and route 2 if $m_1 > \hat{m}$. The artificial objective function $E\left(n_1^e(m_1), m_1\right)$ for the user equilibrium scenario introduced in subsection 2.1 is also shown here as well as function Γ introduced in subsection 3.1 and showing the total cost under condition that the cars choose their routes so that the total travel cost is minimized. Function Γ coincides with TC_t outside the interval $\left[\breve{m} - N/\delta, \breve{m}\right]$ where Γ is concave.

5.2. The interior solutions

Since $0 < n_1 < N$ only for m_1 in the interval $]\hat{m} - N, \hat{m}[$, where the objectives TC_v, **and** TC_t are affine with slope Ψ,

$$\Psi = 0 \tag{29}$$

is the necessary condition for existence of an interior solution in both PS and SB scenarios. If there is a solution in the interior of the rectangle $[0, N] \times [0, M]$ then there is a whole straight segment of solutions within the rectangle including the two points on the edges. Otherwise, there can be one or two solutions on the edges.

For interior solution, $[\hat{m} - N, \hat{m}]$ also has to overlap with the open interval $]0, M[$, i.e. inequality

$$0 < \hat{m} < N + M \tag{30}$$

has to be satisfied. This implies that existence of interior solution in the PS scenario or in the SB scenario is sufficient for existence of interior solution in the user equilibrium scenario.

However conditions (14) and (15) are not sufficient for existence of an interior solution in the PS or in SB scenario since $TC_l\left(n_1(m_1), m_1\right)$ and $TC_t\left(n_1(m_1), m_1\right)$ may take lower values outside the interval $[\hat{m} - N, \hat{m}]$. In order to guarantee existence of interior solutions, one has to impose additional conditions on the sign of the one-side derivatives of the objective at $\hat{m}$ and $\hat{m} - N$, namely, non-positiveness of left derivative at $\hat{m} - N$ if $\hat{m} - N > 0$ and non-negativeness of right

derivative at $\hat{m}$ if $\hat{m} < M$. Such conditions, combined with (29) and (30), are sufficient for existence of interior equilibrium because TC_l and TC_t are convex outside the interval.

For completeness and comparability, the necessary and sufficient conditions of existence of an interior solution for all three scenarios (user equilibrium, PS and SB) are summarised in the three propositions below.

Proposition 3. *Interior user equilibrium exists if and only if the conditions (29) and (30) are fulfilled.*

Proposition 4. *Interior Solution in the PS scenario exists if and only if the conditions (29), (30), and the following two conditions are fulfilled:*

i) $2\left(B_1 + B_2\right)\hat{m} + A_1 - A_2 - B_2\left(N + 2M\right) \geq 0$ *or* $\hat{m} \geq M$ *and*

ii) $2\left(B_1 + B_2\right)\left(\hat{m} - N\right) + A_1 - A_2 + B_1 N - 2B_2 M \leq 0$ *or* $\hat{m} \leq N$.

If an interior solution in the PS scenario exists then an interior user equilibrium exists and the sets of solutions to the two problems coincide.

Proposition 5. *Interior Solution in the SB scenario exists if and only if the conditions (29), (30), and the following two conditions are fulfilled:*

i) $2\left(B_1 + B_2\right)\hat{m} + A_1 - A_2 - B_2\left(N + 2M\right) - b_2 N \geq 0$ *or* $\hat{m} \geq M$ *and*

ii) $2\left(B_1 + B_2\right)\left(\hat{m} - N\right) + A_1 - A_2 + B_1 N - 2B_2 M + b_1 N \leq 0$ *or* $\hat{m} \leq N$.

If an interior solution in the SB *scenario exists then an interior solution in the PS scenario and an interior user equilibrium exist and the sets of solutions to the three problems coincide.*

5.3. Location of the Second Best equilibria for small number of trucks

In this subsection we investigate behaviour of solutions to the two Stackelberg scenarios when the number M of trucks is small. We compare the results between the two scenarios and with the User Equilibrium and System optimum scenarios considered in the previous sections. Note that when there are no trucks in the Stackelberg scenarios then the cars are in user equilibrium.

Assume first inequality $-b_2 N < a_2 - a_1 < b_1 N$ that is the cars use both routes when there are no trucks. Then, due to continuity of the route cost for cars, the cars continue using both routes when the total number of trucks is small. Then, due to (28), the number of trucks on route 1 is $n_1\left(m_1\right) = \hat{m}\left(M\right) - m_1$, and the route cost difference for trucks is $C_{l,1}\left(\hat{m}\left(M\right) - m_1, m_1\right) - C_{l,2}\left(N - \hat{m}\left(M\right) + m_1, M - m_1\right) = \Psi\left(M\right)$. The purpose of the truck agency in the PS scenario is to minimize the truck total cost $TC_l\left(\hat{m}\left(M\right) - m_1, m_1\right) = m_1\left[A_1 + B_1\hat{m}\left(M\right)\right] + \left(M - m_1\right)\left[A_2 + B_2\left(N + M - \hat{m}\left(M\right)\right)\right]$, which is affine by m_1 with the slope $\Psi\left(M\right)$. The total cost for cars does not depend on m_1 , hence the total cost is also affine with slope $\Psi\left(M\right)$ .Therefore the solution for both PS and SB is $m_1 = 0, n_1 = \hat{m}\left(M\right)$ if $\Psi\left(M\right) > 0$ and $m_1 = M, n_1 = \hat{m}\left(M\right) - M$ if $\Psi\left(M\right) < 0$. Since $\Psi\left(M\right)$ is continuous by M , the location of the solution is determined by the sign of $\Psi\left(0\right)$ as in the User Equilibrium scenario.

Assume now that inequality $a_2 - a_1 > b_1 N$ is fulfilled whence all cars use route 1 when there are no trucks in the system. Then this is true even when there are few trucks, i.e. $n_1 = N$. Substitution into

the objective function for PS gives
$TC_l\left(N,m_1\right)=m_1\left[A_1+B_1\left(N+m_1\right)\right]+\left(M-m_1\right)\left[A_2+B_2\left(M-m_1\right)\right]$ and the optimal solution is $m_1^p=\max\left[0,\min\left(\frac{A_2-A_1-B_1N+2B_2M}{2\left(B_1+B_2\right)},M\right)\right]$. For small M this is equal to 0 if $A_2-A_1-B_1N<0$ and to M if $A_2-A_1-B_1N>0$.

For SB, the objective is $TC_t\left(N,m_1\right)=m_1\left[A_1+B_1\left(N+m_1\right)\right]+\left(M-m_1\right)\left[A_2+B_2\left(M-m_1\right)\right]+N\left[a_1+b_1\left(N+m_1\right)\right]$ and the optimal solution is given by $m_1^s=\max\left[0,\min\left(\frac{A_2-A_1-\left(b_1+B_1\right)N+2B_2M}{2\left(B_1+B_2\right)},M\right)\right]$, which for small M reduces to 0 if $A_2-A_1-\left(b_1+B_1\right)N<0$ and to M if $A_2-A_1-\left(b_1+B_1\right)N>0$.

Summarising, the solution for PS with small M is exactly the same as for User Equilibrium (section 3.2). This is because with small number of trucks the partial marginal cost for trucks is essentially the same as average costs. Table 1 presented for User Equilibrium in Subsection 3.2 is therefore valid for scenario PS as well. When it regards the SB, the solution for the case with cars using different routes in the situation without trucks is the same as PS and User Equilibrium, because the cars can adjust their route choice in such a way that their cost is independent on what route the trucks take. However, the cars have no such ability if they all are on the same route. Table 3 presents all cases and the total cost expressions for the SB scenario with small number of trucks.

Table 3. Solutions to the SB scenario and expressions for total cost when the number of trucks is small.

Conditions on parameters	Solution	Total cost
$-b_2N<a_2-a_1<b_1N$ $\Psi\left(0\right)>0$	$n_1^e=\hat{m}\left(M\right)$, $m_1^e=0$	$\hat{m}\left(a_1+b_1\hat{m}\right)+\left(N-\hat{m}\right)\left[a_2+b_2\left(N+M-\hat{m}\right)\right]+$ $+M\left[A_2+B_2\left(N+M-\hat{m}\right)\right]$
$-b_2N<a_2-a_1<b_1N$ $\Psi\left(0\right)<0$	$n_1^e=\hat{m}\left(M\right)-M$, $m_1^e=M$	$\left(\hat{m}-M\right)\left(a_1+b_1\hat{m}\right)+M\left(A_1+B_1\hat{m}\right)+$ $+\left(N+M-\hat{m}\right)\left[a_2+b_2\left(N+M-\hat{m}\right)\right]$
$a_2-a_1>b_1N$, $A_2-A_1>\left(b_1+B_1\right)N$	$n_1^e=N$, $m_1^e=M$	$N\left[a_1+b_1\left(N+M\right)\right]+M\left[A_1+B_1\left(N+M\right)\right]$
$a_2-a_1>b_1N$, $A_2-A_1<\left(b_1+B_1\right)N$	$n_1^e=N$, $m_1^e=0$	$N\left(a_1+b_1N\right)+M\left(A_2+B_2M\right)$
$a_2-a_1<-b_2N$, $A_2-A_1>-\left(b_2+B_2\right)N$	$n_1^e=0$, $m_1^e=M$	$N\left(a_2+b_2N\right)+M\left(A_1+B_1M\right)$
$a_2-a_1<-b_2N$, $A_2-A_1<-\left(b_2+B_2\right)N$	$n_1^e=0$, $m_1^e=0$	$N\left[a_2+b_2\left(N+M\right)\right]+M\left[A_2+B_2\left(N+M\right)\right]$

Thus the coordination of trucks by government agency can bring a benefit even with a small number of trucks if all cars take the same route. Assume that the marginal cost for trucks on this route is

higher but the average cost is lower than the cost for the empty route. Then the selfish trucks or the trucks controlled by the private agency would take the route with cars but the trucks controlled by government agency would take another route. The latter solution results in higher cost for trucks but lower cost for cars and for all vehicles in total.

5.4. Location of the Second Best equilibria and the benefit of coordination of truck routes with large number of trucks

Similar to the User Equilibrium and System Optimum scenarios, the sign of Δ defined by (14) determines behaviour of the solutions and the expressions for total cost for large values of M as summarized in the following two propositions. First the PS scenario is considered.

Proposition 6. *If* $\Delta > 0$ *then for*

$$M > \max\left[\frac{A_1 - A_2 + B_1 N}{2B_2}, \frac{(b_1+b_2)(A_2-A_1)-(B_1+B_2)(a_2-a_1)}{\Delta B_1 B_2} - N, \right.$$
$$\left. \frac{a_2 - a_1 + b_2 N}{b_1}, \frac{(b_1+b_2)[A_2 - A_1 + B_2 N] - 2(B_1+B_2)(a_2 - a_1 - b_1 N)}{2\Delta B_1 B_2} \right]$$

the solution in the PS *scenario is unique and involves partial separation with all cars on route 1 and trucks on both routes,* $\left(n_1^p, m_1^p\right) = \left(N, \frac{A_2 - A_1 - B_1 N + 2B_2 M}{2(B_1+B_2)} \right)$*, and the total cost*

$$TC_t^p = N(a_1 + b_1 N) + M(A_2 + B_2 M) - \frac{\left[A_2 - A_1 - (b_1 + B_1)N + 2B_2 M\right]^2}{4(B_1+B_2)} + \frac{(b_1 N)^2}{4(B_1+B_2)}. \quad (31)$$

If $\Delta < 0$ *then for*

$$M > \max\left[\frac{A_2 - A_1 + B_2 N}{2B_1}, \frac{(b_1+b_2)(A_2-A_1)-(B_1+B_2)(a_2-a_1)}{\Delta B_1 B_2} - N, \right.$$
$$\left. \frac{a_1 - a_2 + b_1 N}{b_2}, \frac{(b_1+b_2)[A_2 - A_1 - B_1 N] - 2(B_1+B_2)(a_2 - a_1 + b_2 N)}{2\Delta B_1 B_2} \right]$$

the solution in the PS scenario is unique and involves partial separation with all cars on route 2 and trucks on both routes, $\left(n_1^p, m_1^p\right) = \left(0, \frac{A_2 - A_1 + B_2 N + 2B_2 M}{2(B_1+B_2)} \right)$*, and the total cost*

$$TC_t^p = N(a_2 + b_2(N+M)) + M(A_2 + B_2(N+M)) -$$
$$- \frac{\left[A_2 - A_1 + (b_2 + B_2)N + 2B_2 M\right]^2}{4(B_1+B_2)} + \frac{(b_2 N)^2}{4(B_1+B_2)} \quad . \quad (32)$$

Proof. See Appendix A.

Comparison of expressions (31) and (32) with (50) and (51) shows that for large M and $\Delta > 0$ $m_1^c = m_1^o + \frac{b_1 N}{2(B_1+B_2)}$ and $TC_t^c = TC_t^o + \frac{(b_1 N)^2}{4(B_1+B_2)}$ while for large M and $\Delta < 0$ $m_1^c = m_1^o - \frac{b_2 N}{2(B_1+B_2)}$ and $TC_t^c = TC_t^o + \frac{(b_2 N)^2}{4(B_1+B_2)}$. Thus in both cases the trucks overuse the unique route used by the cars by a number independent on the total number of trucks. This occurs

because in the PS scenario the truck company does not care about the congestion costs it incurs upon the cars.

Comparison with equations (26) and (27) demonstrates that PS may magnify or reduce the travel costs compared to the user equilibrium. In particular, for large M and $\Delta > 0$, the change in total travel cost due to the collusion of trucks is equal to $TC_t^c - TC_t^e = \frac{(b_1 N)^2 - \left[A_2 - A_1 + (b_1 - B_1) N\right]^2}{4(B_1 + B_2)}$.

Consider now what happens if the trucks are controlled by a governmental agency with the purpose to reduce the total travel cost.

Proposition 7. *If $\Delta \neq 0$ then for M large enough*[97] *the solution in the SB scenario is unique and coincides with the system optimum.*

Proof. See Appendix A.

Thus when the number of trucks is large enough the governmental agency can enforce the system optimum route choice by all vehicles. According to Proposition 2 this solution involves myopic route choice by car drivers.

5.5. Evolution of the solutions to the Second Best scenarios with increasing number of trucks

Consider now what happens when the number of trucks increases from 0. Similar to subsection 3.4, assume that with $M = 0$ there are cars on both routes, i.e.

Assumption 1. When there are no trucks, the cars use both routes in the user equilibrium, i.e. condition (4) is fulfilled: $-b_2 N < a_2 - a_1 < b_1 N$.

Then for small M, i.e. $M < \min\left[\frac{a_2 - a_1 + b_2 N}{b_1}, \frac{a_1 - a_2 + b_1 N}{b_2}\right]$, we have $\hat{m}(M) - N < 0$ and $\hat{m}(M) > M$ which implies that $TC_l\left(n_1(m_1), m_1\right)$ and $TC_t\left(n_1(m_1), m_1\right)$ are both linear with equal slope $\Psi(M)$ for all $0 \le m_1 \le M$. Then both TC_l and TC_t are minimised at $m_1 = 0$ (no trucks on route 1) if $\Psi(M) > 0$ and at $m_1 = M$ (no trucks on route 2) if $\Psi(M) < 0$. This is the same solution as the equilibrium.

Thus, if there are so few trucks that they cannot crowd out all cars to the same route then the coordination of trucks can reduce neither the total truck cost nor the total cost for all vehicles.

Similar to subsection 2.3.1 and additional to Assumption 1, assume now

Assumption 2. When there are no trucks and the cars choose routes according to the user equilibrium, route 2 is cheaper for trucks than route 1.

Algebraically, Assumption 2 can be expressed as $C_{l,2}\left(n_1(0), 0\right) < C_{l,1}\left(N - n_1(0), 0\right)$ or as $\Psi(M) > 0$.

The role of Assumption 2 in the scenarios with coordinated trucks is the same as for user equilibrium scenario. Assumptions 1 and 2 together imply that for small total number of trucks M all trucks take route 2. As M increases, the new trucks accumulate on route 2 and increase the cost for cars on this route, whence the cars move to route 1.

[97] A lower bound for M is specified in the proof.

Evolution of the solution in the Second Best scenarios is partially similar to the User Equilibrium. With $\Delta > 0$, the solution moves right along the bottom edge and up along the right edge, while with $\Delta < 0$, a jump eventually occurs to the top or to the left edge of the rectangle. However, since the objectives $TC_l\left(n_1(m_1), m_1\right)$ and $TC_t\left(n_1(m_1), m_1\right)$ are not convex in general, jumps do not need to occur when an interior solution exists but may also occur at other values of M.

An example with multiple jumps is demonstrated in Figure 7. Truck usage of route 1 in PS and in SB scenarios are plotted against the total flow of trucks. The plot also shows the benefit of nationalisation of the company controlling the routes of the car. This benefit is equal to the total cost in the PS scenario minus the total cost in the SB scenario. For small values of M, both scenarios assign all trucks to the second route. At $M = 0.34$ there is an interior solution, and all trucks switch to route 1 in both scenarios. In the SB scenario, all trucks switch back to route 2 at $M = 0.56$, start to gradually move to route 1 at $M = 3.53$ and abruptly all switch to route 1 at $M = 3.76$, while in the PS they stay all on route 1. Notice that the benefit of nationalization does not jump but evolves continuously. The jump in route usage does not change the total travel cost.

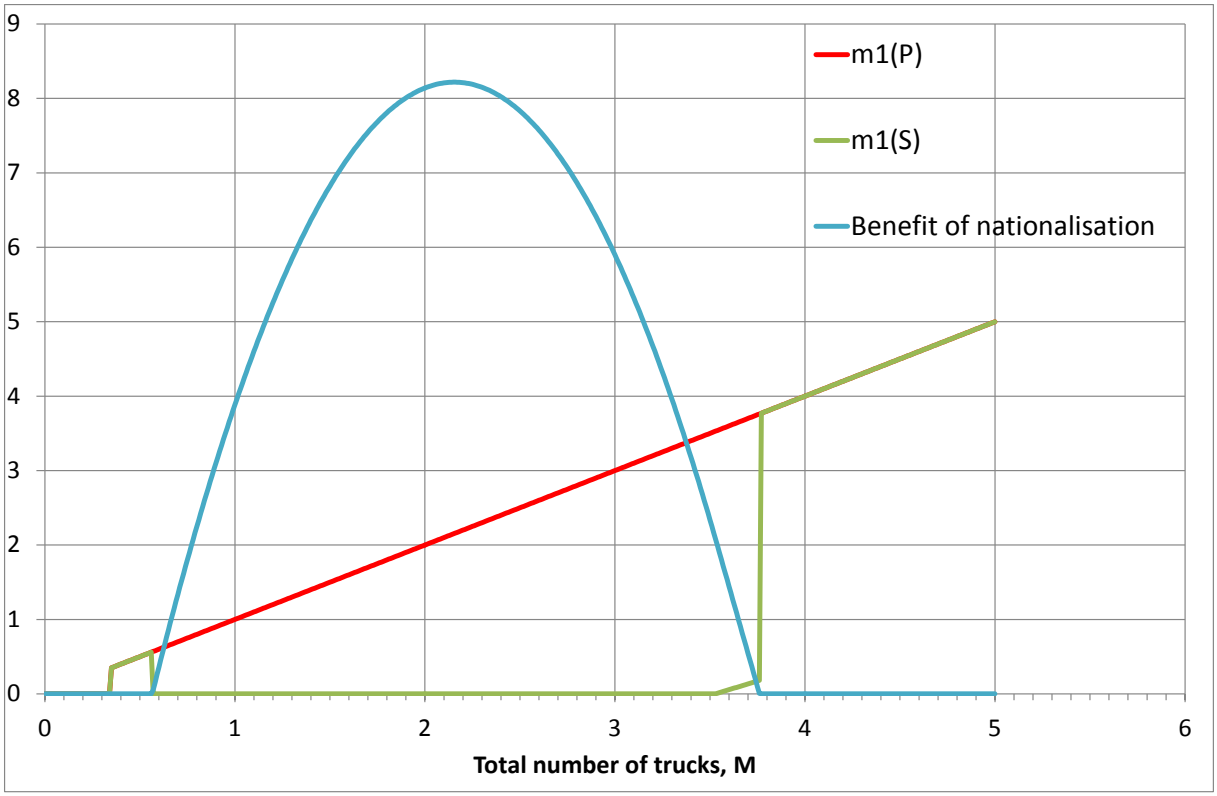


Figure 7. Truck usage of route 1 in scenarios PS and SB, and the Benefit of nationalization vs the total number of trucks. The assumed parameter values are
$a_1 = 1.2,\ a_2 = 25.65,\ b_1 = 1,\ b_2 = 1,\ A_1 = 0.51,\ A_2 = 26.2,\ B_1 = 1.1,\ B_2 = 3.8,\ N = 25$

6. Numerical example with best-guess parameters

In this section we present route usage by cars and by trucks and the total travel cost for all four scenarios -- User equilibrium, System Optimum, PS and SB -- as a function of the total number of trucks with all other parameters fixed. To present a realistic example with the two vehicle classes, we assume that the drivers can choose between a 30 km long urban arterial and a 37,5 km long freeway respectively. The free flow speed of 60 km/h on the arterial and 90 km/h on the freeway are assumed for cars and imply the free flow times 30 min. on the arterial route and 25 min. on the freeway route. Assume the total car flow 1000 vehicles/h. The arterial road is also assumed to be

more sensitive to the congestion, with congestion coefficient 0,1 min/veh compared to 0,05 min/veh for the freeway. Without any trucks, the volume of cars on the arterial route would be $n_1^e(0) = 300$ vehicles in the User Equilibrium scenario and $n_1^o(0) = 317$ vehicles in the System Optimum scenario.

For the trucks, we assume lower free flow speed than for cars: 50 km/h on the arterial and 80 km/h on the freeway, which corresponds to the free flow travel times 36 min. and 28 min. When the volume reaches such level that the time for cars doubles compared to the free flow situation, overtaking the trucks becomes problematic due to the high congestion. Let us call it the capacity volume. From the assumed parameters, the capacity can be calculated as 300 on the first route and 500 on the second route. From that, the congestion coefficients for trucks are estimated to 0,08 on the arterial and 0,44 on the freeway, so that, at volume equal to the capacity, the truck travel time equals the car travel time.

To complete the specification, we need the passenger car equivalent (PCE) factor for trucks and the values of time. Florian and He (2005) present a table with PCE factors for various traffic mixes, truck sizes, slopes and link lengths. From that table, we pick a value 3,9 corresponding to a medium heavy truck, the slope 0-2 grade, the proportion 5-10% trucks on the road, and a road link longer than 1 km. The free flow time for trucks and the congestion coefficient for truck have to be divided by this value to obtain the travel time expenditure per PCE.

There is no consensus regarding the values of time for commercial vehicles. However most of literature seems to agree that that value is 2 to 5 times higher than for cars and that the average value of time for passenger cars is around 10 euro per hour. We assume 10 euro per hour for cars and 40 euro per hour for trucks.

Thus the model parameters in Euro and in Euro per PCE for our example are $a_1 = 5.0, a_2 = 4.16, b_1 = 0.016, b_2 = 0.0083, A_1 = 6.15, A_2 = 4.81, B_1 = 0.014, B_2 = 0.0075$, and N=1000 vehicles. For each scenario, the resulting car volumes and truck volumes on route 1 and the welfare loss (in Euro per hour) compared to the Social Optimum scenario are presented in Figure 8 as functions of the total flow of trucks (PCE per hour).

When the number of trucks increases from 0, they first all accumulate on route 2 and gradually crowd out the cars to route 1. This is common for all four scenarios. The total cost difference between the Social Optimum and all other scenarios comes from the small difference in allocation of cars. With increasing congestion, this difference becomes even smaller, because all scenarios have solutions very close to the System Optimum. However, at $M = 294$ it becomes more beneficial for the society to put all trucks and just 138 cars on route 1. The other scenarios do not realise this possibility because the average cost for cars is still lower on route 1 than on route 2. So the second best strategies do not improve the situation compared to the User equilibrium, when the number of trucks is small. From $M = 294$ the travel cost gain increases steeply and reaches 40 Euro per hour at $M = 461$ where the jump occurs in the SB scenario. This jump brings a complete separation of the vehicle groups: all trucks on route 1 and all cars on route 2. Note that this jump is not connected to the interior solution. Actually, no interior solution for SB exists in our example for any value of M .

After the jump, the situation in SB is close to the System Optimum. From that stage, the welfare loss for SB diminishes and, starting from $M = 521$, SB coincides with System Optimum. Thus $M = 521$ is in our example large enough for the coordinator of the truck fleet to incur the Social Optimum.

For $M > 461$, the User Equilibrium and the PS still have all trucks on route 2 until $M = 500$ where the travel cost for trucks coincides on two routes, and a set of interior solutions emerge. The solution in these two scenarios jumps to the situation where all cars are on route 2 and almost all trucks are on route 1. From $M = 529$, the welfare loss of User Equilibrium stabilises at 57 Euro but the welfare loss of PS continues to increase until $M = 7675$ where it stabilizes at 821 Euro per hour. Thus in our

example the coordination of trucks by a private firm does not have any effect unless the total flow of trucks is very large, in which case it has a large negative welfare effect.

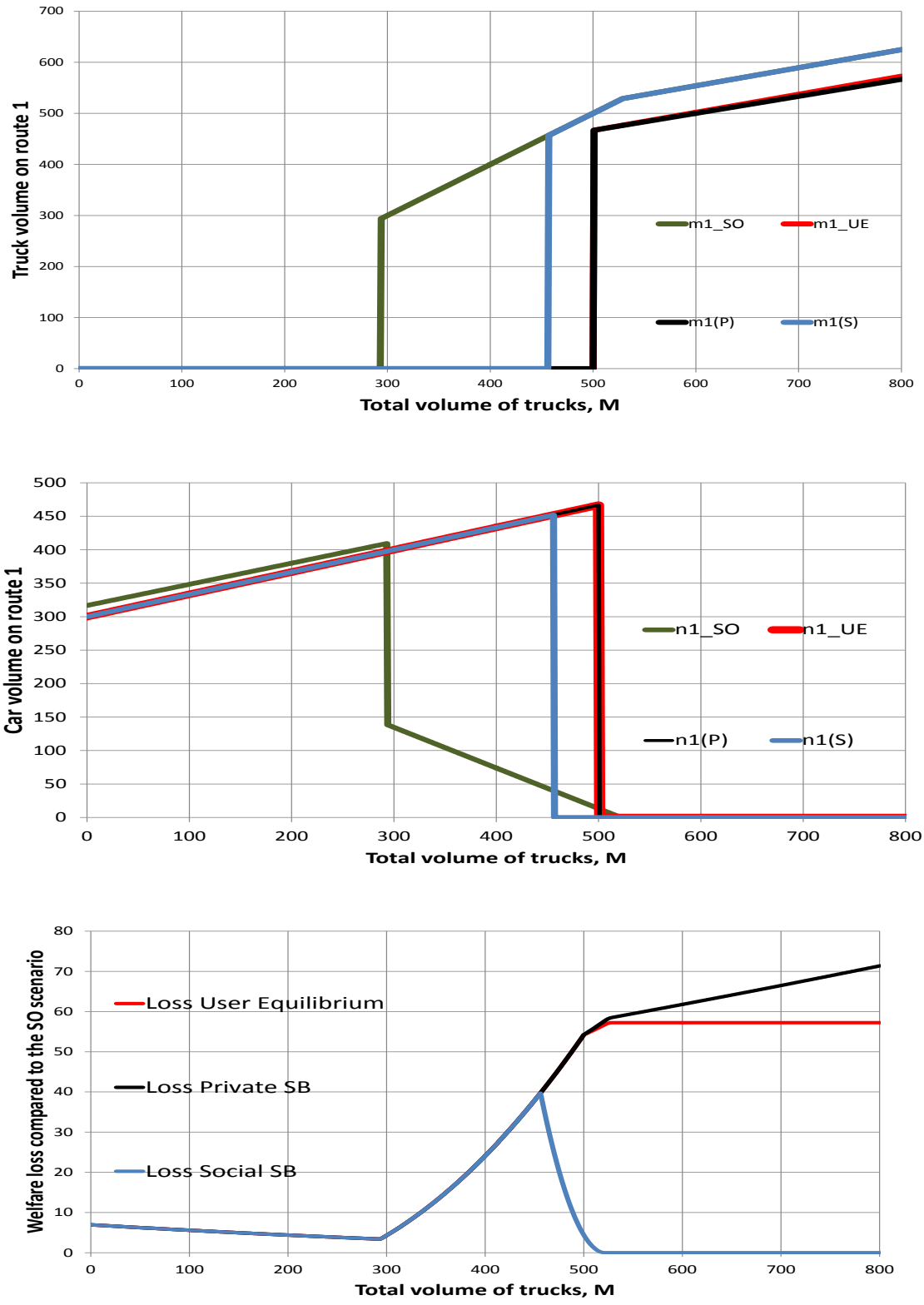


Figure 8. Car volumes and truck volumes on route 1 in the four scenarios and the welfare loss compared to the Social Optimum scenario, for the parameter values N=1000,
$a_1 = 5.0, a_2 = 4.16, b_1 = 0.016, b_2 = 0.0083, A_1 = 6.15, A_2 = 4.81, B_1 = 0.014, B_2 = 0.0075$.

7. Conclusions

This paper has shown that coordination of a part of the vehicles has a potential for social saving if the number of coordinated vehicles is large enough. It also shows that the aim of the coordination has crucial importance for the outcome of the game. The type of coordination envisaged matters. In this sense, there is a need to envisage the coordination issues studied in this paper in the perspective of mechanism design problem.

The practical implementation of the coordination mechanisms needs to be examined. Some agent may have high coordination costs, may be reluctant to centralized planning of their operation, or may just be unaware of the potential savings of coordination. We know from several studies of road pricing that the acceptability issues and the costs of implementation are likely to play a crucial role. In the problem studied here, the role of the local authorities and of the government will be essential to put ahead the right incentives. Such incentive can be reduced tolls or access to parking or to special lanes for the fleets which agree to participate in specific coordination schemes.

Finally, the model developed here is based on a static analysis of congestion. That is the level of congestion is assumed to be constant over the peak period. The proposed model relies on the simplest setting (two routes in parallel, static congestion and linear congestion function). Obviously, these are restrictive hypothesis. In particular, a dynamic model, that is a model where the level of congestion depends on the time of the day, will allow a much richer analysis. This is because coordination over the time of the day is potentially much more fruitful than the spatial coordination explored in this paper. Temporal coordination is used at a large scale, for example, to organize holiday. Temporal regulations are effective in reducing the noise during the nights and the weekends (for example, it is used for airplanes). Our approach suggests that large enough operators may discipline themselves in order to better use the infrastructure over the time of the day. Preliminary results suggest that the social savings of such coordination are likely to be substantial.

References

Arnott, R., de Palma, A. and Lindsey, R. (1993) A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand. The American Economic Review, 83(1), 161 - 79.

Brueckner, J. K. (2002) Airport Congestion When Carriers Have Market Power. The American Economic Review, 92(5), 1357-1375.

Chi Kin Chau, C. K. and Sim K.M. (2003) The price of anarchy for non-atomic congestion games with symmetric cost maps and elastic demands. Operations Research Letters 31, 327–334.

Florian, M. and He, S. (2005) Assigning mixed traffic of cars and trucks: how to handle asymmetric interactions. Presented at Ontario EMME/2 Users' Group Meeting, Toronto, April 2005

Harks, T. (2011) Stackelberg Strategies and Collusion in Network Games with Splittable Flow. Theory Comput. Syst. 48, 781–802.

Korilis, Y.A., Lazar, A.A., Orda, A. (1997) Achieving network optima using Stackelberg routing strategies. IEEE/ACM Trans. Netw. 5(1), 161–173.

Koutsoupias, E., Papadimitriou, C.H. (1999) Worst-case equilibria. In: Proc. of the 16th Annual Symposium on Theoretical Aspects of Computer Science (STACS). Lecture Notes in Computer Science, 1563, 404–413.

Perakis, G. (2007) The “Price of Anarchy” Under Nonlinear and Asymmetric Costs. Mathematics of Operations Research, 32(3), 614–628.

Roughgarden, T. (2002) The price of anarchy is independent of the network topology. J. Comput. Syst. Sci. 67, 341–364.

Roughgarden, T. (2004) Stackelberg scheduling strategies. SIAM J. Comput. 33(2), 332–350.

Roughgarden, T., Tardos, E. (2002) How bad is selfish routing? J. ACM 49(2), 236–259.

Appendix C1. Proofs of the propositions.

Proof of Proposition 1. Inequality $M > \frac{A_2 - A_1 + B_2 N}{B_1}$ implies that the trucks cannot be all on route 1, and from inequality $M > \frac{A_1 - A_2 + B_1 N}{B_2}$ follows that they cannot be all on route 2. If $\Delta > 0$, then inequality $M > \frac{(A_2 - A_1)(b_1 + b_2) - (a_2 - a_1)(B_1 + B_2)}{\Delta B_1 B_2} - N$ is equivalent to $\Psi(M) > 0$, therefore the solution is unique. Since the trucks use both routes, we have $C_{l,1}(n_1, m_1) = C_{l,2}(N - n_1, M - m_1)$, i.e.

$$n_1 + m_1 = \frac{A_2 - A_1 + B_2(N + M)}{B_1 + B_2}. \qquad (A.1)$$

But $\Psi(M) > 0$ can also be written as $\frac{A_2 - A_1 + B_2(N + M)}{B_1 + B_2} < \frac{a_2 - a_1 + b_2(N + M)}{b_1 + b_2}$. It follows that $n_1 + m_1 < \frac{a_2 - a_1 + b_2(N + M)}{b_1 + b_2}$, or $C_{v,1}(n_1, m_1) < C_{v,2}(N - n_1, M - m_1)$, which means that all cars must be on route 1. The expressions for m_1^e and for the total cost are obtained now by substitution of $n_1 = N$ into (A.1) and into equation (17).

If $\Delta < 0$, then inequality $M > \frac{(A_2 - A_1)(b_1 + b_2) - (a_2 - a_1)(B_1 + B_2)}{\Delta B_1 B_2} - N$ is equivalent to $\Psi(M) < 0$, and the rest of the proof is performed the similar way.

Proof of Proposition 2. Remember, that for any M, the objective function $\Gamma(m_1, M)$ is continuously differentiable and is a convex quadratic function of m_1 for $m_1 \le \breve{m}(M) - N/\delta$, concave quadratic for $\breve{m}(M) - N/\delta < m_1 < \breve{m}(M)$ and again convex quadratic for $m_1 \ge \breve{m}(M)$. Denote m_o' and m_o'' the abscissae of the vertices of the left and the right convex parabolas, i.e. $m_o' = \frac{A_2 - A_1 - (b_1 + B_1)N + 2B_2 M}{2(B_1 + B_2)}$ and $m_o'' = \frac{A_2 - A_1 + (b_2 + B_2)N + 2B_2 M}{2(B_1 + B_2)}$.

With $\Delta > 0$, the inequalities $M > \frac{A_1 - A_2 + (b_1 + B_1)N}{2B_2}$, $M > \frac{2b_2 N - a_1 + a_2}{b_1 + B_1}$, and $M > \frac{(b_1 + b_2 + B_1 + B_2)\left[A_2 - A_1 + (b_2 + B_2)N\right] + 2(B_1 + B_2)(2b_1 N + a_1 - a_2)}{2\Delta B_1 B_2}$ are equivalent to inequalities $m_o' > 0$, $\breve{m} < M$, and $\breve{m} - N/\delta > m_o''$ respectively. Therefore, $0 < m_o' < m_o'' < \breve{m} - N/\delta < \breve{m} < M$. The objective (as function of m_1) decreases on $[0, m_o']$, increases on $[m_o', \breve{m} - N/\delta]$, increases on $[\breve{m}, M]$ since $m'' < \breve{m}$, and increases on $[\breve{m} - N/\delta, \breve{m}]$

since it is concave on this interval and $\frac{\partial \Gamma(\breve{m}, M)}{\partial m_1} < 0$. Thus the objective has a unique minimum at $m_1 = m_o'$ and the number of cars $n_1 = N$ is determined uniquely by (18).

The proof for the second case $\Delta > 0$ is quite similar.

Proof of Proposition 6. Remind, that for any M, the objective function $TC_l\left(n_1(m_1), m_1\right)$ is continuous and is a convex quadratic function of m_1 for $m_1 \le \hat{m}(M) - N$, affine with slope $\Psi(M)$ for $\hat{m}(M) - N < m_1 < \hat{m}(M)$ and again convex quadratic for $m_1 \ge \hat{m}(M)$. Denote m_p' and m_p'' the abscissae of the vertices of the left and the right convex parabolas, i.e.

$m_p' = \frac{A_2 - A_1 - B_1 N + 2B_2 M}{2(B_1 + B_2)}$ and $m_p'' = \frac{A_2 - A_1 + B_2 N + 2B_2 M}{2(B_1 + B_2)}$.

With $\Delta > 0$, the inequalities $M > \frac{A_1 - A_2 + B_1 N}{2B_2}, M > \frac{a_2 - a_1 + b_2 N}{b_1}$,

$M > \frac{(b_1 + b_2)\left[A_2 - A_1 + B_2 N\right] - 2(B_1 + B_2)(a_2 - a_1 - b_1 N)}{2\Delta B_1 B_2}$, and

$M > \frac{(b_1 + b_2)(A_2 - A_1) - (B_1 + B_2)(a_2 - a_1)}{\Delta B_1 B_2} - N$ are equivalent to inequalities $m_p' > 0$, $\hat{m} < M$, $\hat{m} - N > m_p''$, and $\Psi(M) > 0$ respectively. Therefore, $0 < m_p' < m_p'' < \hat{m} - N < \hat{m} < M$. The objective (as function of m_1) decreases on $\left[0, m_p'\right]$, increases on $\left[m_p', \hat{m} - N\right]$, increases on $\left[\hat{m}, M\right]$ since $m'' < \hat{m}$, and increases on $\left[\hat{m} - N, \hat{m}\right]$ since $\Psi(M) > 0$. Thus the objective has a unique minimum at $m_1 = m_p'$ and the number of cars $n_1 = N$ is determined uniquely by (28).

The proof for the second case $\Delta > 0$ is quite similar.

Proof of Proposition 7. Assume $\Delta > 0$. Then for M satisfying the condition of Proposition 2 the system optimum is unique and attained at $(n_1, m_1) = (N, m_o')$. If we additionally require that

$$M > \max\left[\frac{A_1 - A_2 + (b_1 + B_1)N}{2B_2}, \frac{(b_1 + b_2)(A_2 - A_1) - (B_1 + B_2)(a_2 - a_1)}{\Delta B_1 B_2} - N,\right.$$
$$\left.\frac{a_2 - a_1 + b_2 N}{b_1}, \frac{(b_1 + b_2)\left[A_2 - A_1 + (b_2 + B_2)N\right] - 2(B_1 + B_2)(a_2 - a_1 - b_1 N)}{2\Delta B_1 B_2}\right]$$

then, similar to the proof of Proposition 6, it can be shown that solution to the SB scenario is unique and also equal to (N, m_o'). A similar reasoning applies to the case $\Delta < 0$.

Annex D: Comparison of two dynamic transportation models: The case of Stockholm congestion charging

Refrence:

de Palma, A., L. Engelson, I. Kristoffersson, M. Saifuzzaman and K. Motamedi (2012), Comparison of two dynamic transportation models: The case of Stockholm congestion charging. In *proceedings of the 4th TRB Conference on Innovations in Travel Modeling*, Florida, USA.

Abstract

This paper reviews the transportation models used for predicting impacts of congestion charging in European cities and carries out in-depth comparison of two such models, METROPOLIS and SILVESTER. Both are mesoscopic dynamic models involving modal split, route choice and departure time choice calibrated for the Stockholm baseline situation without charges and applied for modeling effects of congestion charging. The results obtained from the two models are mutually compared and validated against actual outcome of the Stockholm congestion charging scheme. Both models provide significant improvement in realism over static models. However results of cost benefit analysis may differ substantially.

1. Introduction

There is a consensus that congestion charging in combination with other congestion mitigation measures is a proper instrument for reducing the adverse impacts of transportation on environment and improving citizens' quality of life. The interest towards design of effective congestion charging systems is growing in many countries and especially in large cities where congestion has become a burning issue. The transportation planning professionals agree that travel forecasts using a good quality regional transportation model is necessary for design of the charging system as well as for evaluation of a system in use.

There is a large scientific literature available on impacts of congestion charging (Pigou, 1920; Vickrey, 1969; Small, 1983; Arnott et al., 1994; Glazer and Niskanen, 2000). The literature considering modeling of congestion charging is however more limited, e.g. Koh and Shepherd (2006). In practice, static assignment models integrated with travel demand models are often applied to forecast the impact in feasibility studies of congestion charging. This has been the case for example in Oslo (Odeck et al., 2003), Stockholm (Eliasson and Mattsson, 2006) and Copenhagen (Rich and Nielsen, 2007; Nielsen et al., 2002). It has however been agreed in the research community that the temporal aspects of congestion have a crucial role on system level. For example, the forecasts made with static models for Stockholm congestion charging system resulted in severe overestimation of impact on traffic flows during the peak hour and, at the same time, great underestimation of changes in travel times (Engelson and van Amelsfort, 2011). Moreover, the most effective charges aim to redistribute trips in time in order to cut down the congestion peak. Therefore impact of time-varying charges on departure time choice is an important issue. A mesoscopic dynamic model (MDM) can capture the time-varying aspect of congestion and congestion charging. At the same time it is not as detailed as a microscopic model. A mesoscopic assignment model integrated with a travel demand model is therefore suitable for calibration of whole city networks and thus for modeling impacts of city-wide congestion charging schemes. For a recent survey of dynamic models we refer the reader to de Palma and Fosgerau (2011).

It is however not obvious which properties of the MDM that is most important for predicting impacts of congestion charging. The aim of this paper is therefore to compare the predictive capability of two MDMs in order to find properties important for correct prediction of congestion charging. METROPOLIS (de Palma et al., 1997) and SILVESTER (Kristoffersson and Engelson, 2009) are two state-of-the-art MDMs developed in the last decade with specific focus on congestion charging applications. De Palma et al. (2005) analyze different congestion charging schemes using METROPOLIS and a stylized urban road network. Marchal and de Palma (2001) apply METROPOLIS to Paris, and also give guidelines for model designers and planners who consider a shift to dynamic traffic simulation. Using METROPOLIS de Palma and Lindsey (2006) assess phase implementation of charging in Paris. SILVESTER is applied to Stockholm in Kristoffersson (2011). Kristoffersson and Engelson (2011) use SILVESTER to evaluate efficiency and equity of alternative congestion charging schemes for Stockholm.

There are very few opportunities to validate transportation models by observed response to charging. In Stockholm we have the unique possibility to use measurements from the field to validate transport models. Therefore both SILVESTER and METROPOLIS are in this paper calibrated to Stockholm conditions in the situation without charging. Model response to the charges are then compared both between the two transport models and to measurements; this in order to provide a benchmark for modeling of congestion charging and in order to find model properties that are important for correct prediction. A similar in-depth comparative study of transportation models suitable for predicting impacts of congestion charging has to our knowledge not been undertaken before. Given that METROPOLIS and SILVESTER share the same ambition to improve conventional static transportation modeling of impacts of congestion charging by using dynamic modeling, but approaches the task in different ways, there is a good opportunity to compare implications of different modeling strategies.

The structures of the two models are described in the next section, followed by a section on how the models have been estimated and calibrated for Stockholm conditions. Section 4 discusses results of the model comparison and Section 5 concludes.

2. Brief description of METROPOLIS and SILVESTER

METROPOLIS is a traffic planning model which uses event based dynamic simulation. It was developed in Geneva by André de Palma, Fabrice Marchal and Yurii Nesterov (de Palma et al., 1997) and later on applied at the University of Cergy-Pontoise by de Palma and Marchal (de Palma and Marchal, 2002). METROPOLIS is based on a simple economic principle, explained originally in Vickrey (1969) and Arnott, de Palma and Lindsey (1993). SILVESTER is also a traffic planning model which uses dynamic simulation. SILVESTER has been developed at KTH Royal Institute of Technology in Stockholm by Leonid Engelson and Ida Kristoffersson (Kristoffersson and Engelson, 2009).

METROPOLIS describes the joint mode, departure time and route choice decisions of drivers. Each vehicle is described individually by the simulator. However, the modeling of congestion on links is carried out at the aggregate or macroscopic level. On the supply side a congestion function (bottleneck, BPR or DAVIS) describes the link travel delays. Demand is represented at microscopic level and each trip can be simulated. Users' characteristics which are necessary for modeling are: valuations of cost and travel time, early and late schedule delay parameters, distribution of preferred arrival times (PATs), and mode choice parameters like valuation of travel time for public transport (PT) and PT penalty or fee. In simulation, each trip is followed individually in its choices of mode, departure time and route. The user chooses the mode considering the average maximum expected utility (Logsum) offered by the car network in comparison with other modes. The choice of departure time for PT is not described by the model, since the PT travel times are external inputs to METROPOLIS. The departure time choice model for car is a continuous logit model, where the individual selects the departure time that minimizes the generalized cost function. METROPOLIS uses a model of route choice based on point-to-point dynamic travel times. The user selects the dynamic shortest path from the origin node to the destination node. The decision will be based on the real time situation of the immediate link and memorized information about the rest of the network up to the destination. It should also be noted that one day corresponds to one iteration in METROPOLIS. The software uses a learning process where users acquire knowledge about their travel and use this information to modify their trip for the next day.

SILVESTER includes the same traveler choices as METROPOLIS: mode, departure time and route choice. However, unlike METROPOLIS, the model is built up of two parts: (1) a model for mode and departure time choice and (2) a model for route choice and calculation of route travel times and costs. SILVESTER iterates between these two parts to reach convergence between demand and supply. The mesoscopic dynamic assignment model CONTRAM (Taylor, 2003) calculates route choice and resulting travel times and monetary costs for trips in each OD-pair, given the demand for car trips departing in each fifteen minute interval. In CONTRAM, vehicles are grouped into packets that are routed through the network. Network supply is described in more detail in CONTRAM than in METROPOLIS, with signal plans coded explicitly as well as conflicting flows at intersections. Demand in the form of time sliced OD-matrices is produced by the model for mode and departure time choice and submitted to CONTRAM. However, preference heterogeneity is explicitly represented through a mixed logit model (Börjesson, 2008) for departure time and mode choice (car or public transport), which takes the travel times and costs from the assignment model and generates the demand for car trips departing in each fifteen minute interval. The time discretization into fifteen minute intervals is a difference compared to METROPOLIS in which time is continuous. The SILVESTER mixed logit model for departure time and mode choice needs user characteristics similar to METROPOLIS: cost and time valuations, early and late schedule delay parameters, and mode choice parameters such as travel time valuation and alternative specific constant for PT. However, some differences exist between the demand model specifications. The mixed logit model in SILVESTER includes also travel time

uncertainty as described by the standard deviation of travel time and the PT alternative includes a dummy for season ticket. Furthermore, desired time of travel is given as a distribution of preferred departure times (PDTs) instead of PATs. Table 1 compares the utility functions for mode and departure time choice in METROPOLIS and SILVESTER. In the utility functions *T* is travel time, *M* is monetary cost, *E* is early schedule delay, *L* is late schedule delay, *and σ* is standard deviation of travel time, with index *t* referring to the departure time. Furthermore, *δ* is a dummy for PT season ticket and *ε* is an error term. Parameter values for the Stockholm application will be given in the next section. Similarly to METROPOLIS, PT travel times do not depend on time-of-day and are external inputs to the SILVESTER model. Route choice in SILVESTER is performed by assigning packets to the network in the order of departure time and finding their shortest paths. That later packets can affect the route choice of earlier packets is accounted for by iterations, starting the assignment process over again after going through all packets. Just as in METROPOLIS, these iterations can be seen as corresponding to a learning process.

Table 1: Comparison of utility functions in METROPOLIS and SILVESTER

METROPOLIS (nested logit for mode choice, continuous logit for departure time choice)	SILVESTER (mixed logit for mode choice and for departure time choice)
$U_{ct} = TIME * T_t + COST * M_t +$ $SDE * E_t + SDL * L_t + \varepsilon_t$ $U_p = TIMEP * T_p + CPT + \varepsilon_p$	$U_{ct} = TIME * T_t + COST * M_t + SDE * E_t +$ $SDL * L_t + TTU * \sigma_t + \varepsilon_t$ $U_p = TIMEP * T_p + ST * \delta + CPT + \varepsilon_p$

The output from METROPOLIS and SILVESTER can be both aggregate and disaggregate. Aggregate data includes network measures of efficiency such as average travel time, average speed, collected revenues, consumer surplus, congestion and mileage. Disaggregate data includes traffic flow on some selected links, temporal distribution of flow on selected links and travel time on some road sections.

3. Application of the two models for Stockholm, baseline situation

This section describes how SILVESTER and METROPOLIS have been estimated and calibrated to Stockholm conditions in the baseline situation without congestion charging. By estimation we mean finding the behavioral parameters on the demand side, i.e. parameters of the departure time and mode choice models. This includes estimation of scheduling, time, and cost parameters. Calibration refers to the adjustment of the complete transportation model (both demand and supply side) to match field measurements in the base line situation, which is the situation without congestion charging.

3.1. Estimation and implementation of demand models

The same data is used for estimating the behavioral parameters of both SILVESTER and METROPOLIS. This data consists of stated and revealed preference data from car drivers crossing the bridge "Tranebergsbron" (which lies just outside the city core of Stockholm, in west direction) driving into the CBD on a work day morning between 6 and 10 am (Börjesson, 2006). Data was collected before introduction of charging in Stockholm, but the stated preference data contains responses to an extra monetary cost on driving. Demand models for both SILVESTER and METROPOLIS are estimated using the software Biogeme (Bierlaire, 2003). Three demand models are estimated for SILVESTER/METROPOLIS: (1) *business* trips, (2) work trips with *fixed* schedule and school trips and (3) work trips with *flexible* schedule and other trips.

The estimation of the mixed logit model for SILVESTER is described in more detail in Börjesson (2008). For implementation in SILVESTER the mixed logit model has been re-estimated because the extra scheduling penalty for early departure time periods did not work well in implementation. The

model for mode and departure time choice estimated for METROPOLIS differs from the model implemented in SILVESTER in two relations: First, instead of mixed logit a nested logit model has been estimated for METROPOLIS. Second, scheduling constraints are on the departure side in the SILVESTER model, whereas they are on the arrival side in the METROPOLIS model. See also the previous section for description of similarities and differences between the two models. Tables 2-4 compare the parameters of the demand models for each trip purpose in METROPOLIS and SILVESTER using the specifications of the utility functions described in Table 1. Mode choice is not available for business trips and the PT parameters are therefore not present in the demand model for business trips.

Table 2: Parameters for business trips in METROPOLIS and SILVESTER

Parameter	METROPOLIS	SILVESTER
TIME	-0.0688	-0.1924
COST	-0.0262	-0.1157 (0.1886)[98]
SDE	-0.0339	-0.1426 (0.1280)
SDL	-0.0428	-0.2825 (0.2557)
TTU	-	-0.1083

Table 3: Parameters for *fixed* trips in METROPOLIS and SILVESTER

Parameter	METROPOLIS	SILVESTER
TIME	-0.0124	-0.1862
COST	-0.0145	-0.2160 (0.2319)
SDE	-0.0152[99]	-0.1662 (0.1261)
SDL	-0.0189	-0.2478 (0.1318)
TIMEP	-0.0465	-0.2214
CPT	-1.6404	-0.05
TTU	-	-0.064
ST	-	13.4886
logsum parameter	4.77	-

Table 4: Parameters for *flexible* trips in METROPOLIS and SILVESTER

Parameter	METROPOLIS	SILVESTER
TIME	-0.0494	-0.2439
COST	-0.0372	-0.1921 (0.1558)
SDE	-0.0200	-0.1958 (0.1929)
SDL	-0.0190	-0.2020 (0.1675)
TIMEP	-0.0687	-0.1838
CPT	-4.9416	-1.3500
TTU	-	-0.0629
ST	-	10.8959
logsum parameter	3.9796	-

In SILVESTER, the preferred departure times are distributed on the interval 6:30-9:30 AM and the simulation is performed for the same period. Travelers who choose their departure time outside this

[98]The values given are mean and standard deviation of the draws of the mixed logit model used in simulation

[99]In METROPOLIS early arrival penalty should be lower than value of time i.e.SDE<TIME, in order to obtain convergence in terms of expected and observed travel time. Only SDE for fixed trips does not follow the criteria and therefore, has been modified to 0.012. Standard error of the estimation(0.0087) allows us to do so.

period are recorded but do not affect the travel costs in the next iteration of the demand model. In METROPOLIS, each traveler's experience is used to modify their departure time on next day. The learning module collects travel information's inside the simulation period. Therefore the simulation period needs to be extended in METROPOLIS beyond the concerned period. A simulation period of 5:00-11:00AM was selected and the demand matrix was extended for this period by putting some extra demand on both ends.

3.2. Calibration

SILVESTER is based on CONTRAM model for Stockholm that has been used and calibrated for decades. The signal plans and saturation flows were adapted to correctly represent the actual traffic situation in Stockholm. The link capacities for the before-charges situation in CONTRAM are consistent with saturation flows and conflicting flows at each intersection. These capacities were imported to METROPOLIS and used in the simple bottleneck congestion functions.

Calibration of SILVESTER and METROPOLIS was performed using field measurements from the situation *without* charging. Field data contained flow measurements for 59 validation links in twelve time periods between 6:30-9:30 am. Furthermore, we use field measurements of average travel time between 7:00-9:00 am on 11 road sections. For validation of travel times the travel time data has been collected using a video technique with automatic license plate matching. Figure 1 shows the location of the links with flow counts and the road sections with travel time measurements used for the calibration.

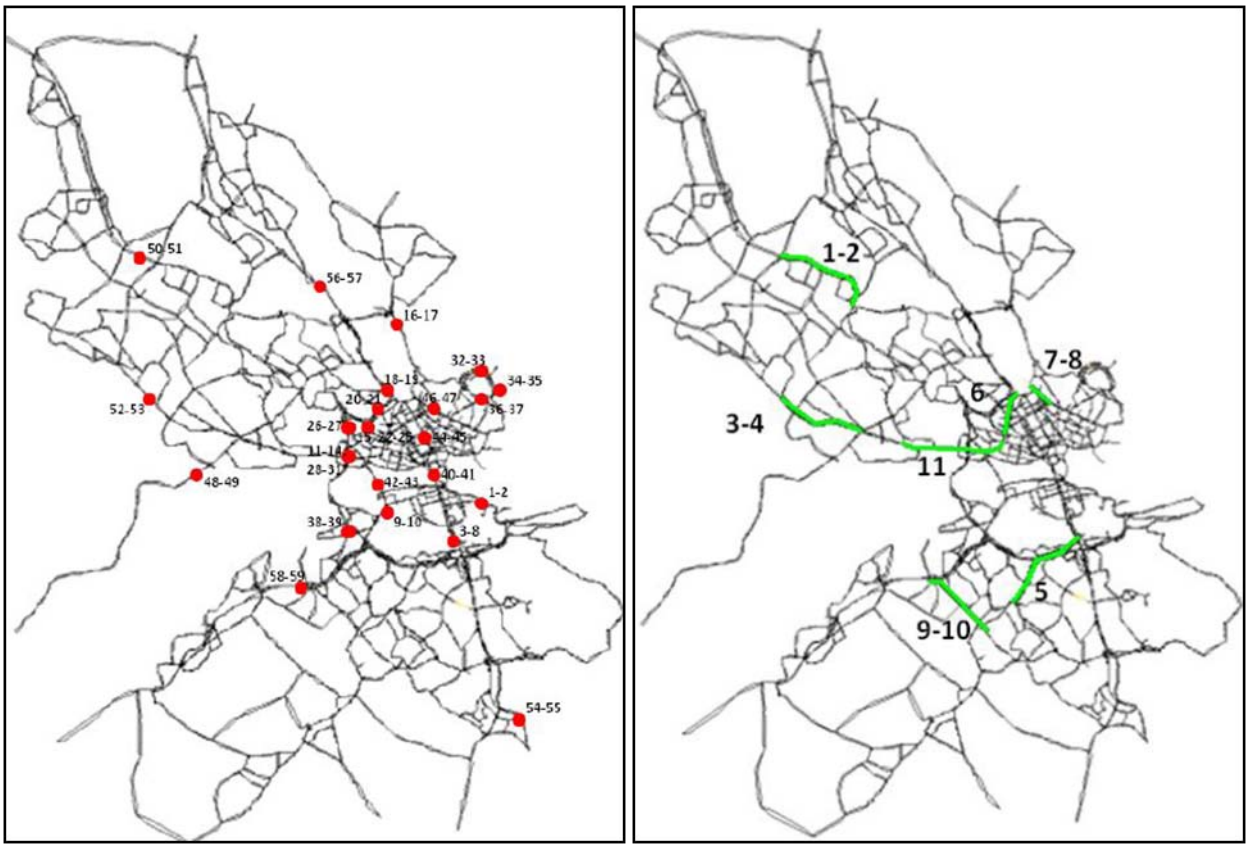


Figure 1: Position of links (to the left) for flow measurement and road sections (to the right) for travel time measurement

The SILVESTER preferred departure times were calibrated using reverse engineering (Kristoffersson and Engelson, 2008). This method takes as input (1) an OD-matrix (calibrated against link flow field measurements) with number of vehicles starting in each actual departure time (ADT) interval and (2) probabilities from the estimated departure time choice model. The demand in each preferred

departure time (PDT) interval is then adjusted such that ADT flow rates are reproduced keeping demand and supply consistent.

The reverse engineering approach is not suitable for METROPOLIS, since the time-sliced demand matrices are not directly available. Instead, the SILVESTER PDT-distributions shifted forward by the free-flow travel times were taken as an initial guess for the PAT-distributions in the METROPOLIS model. These initial PAT-distributions were then calibrated by changing the level of demand and shifting the distributions later. It was done in order to achieve a good fit of simulated link flows to field measurements in the baseline situation. The spillback effect was not considered and simple bottleneck function was used as the congestion function in METROPOLIS.

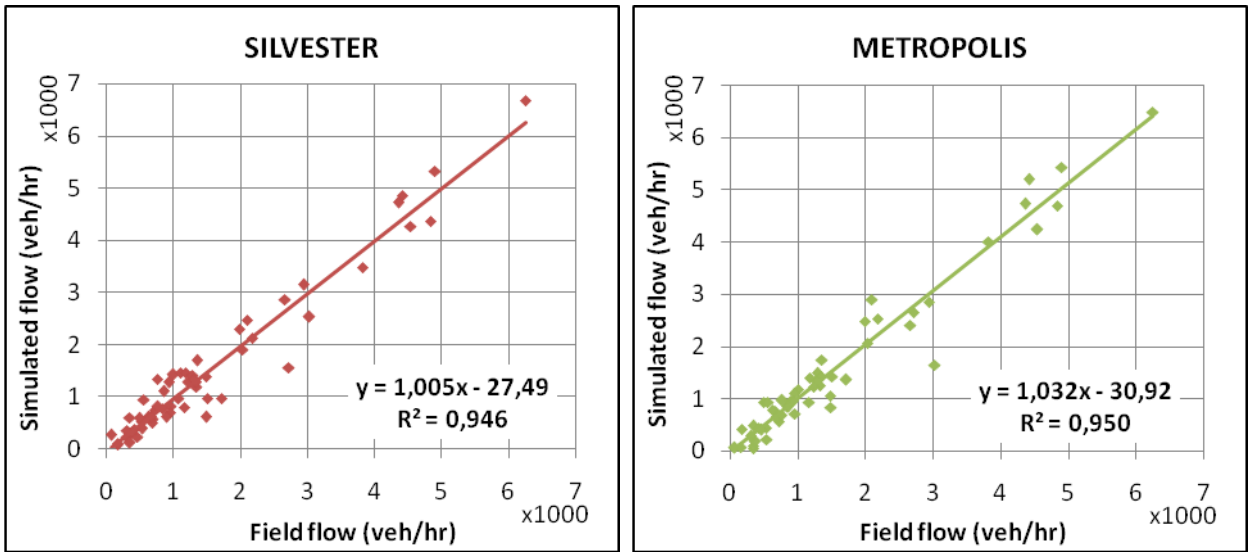


Figure 2: Field vs. Simulated flow in 59 calibration links for before charging situation

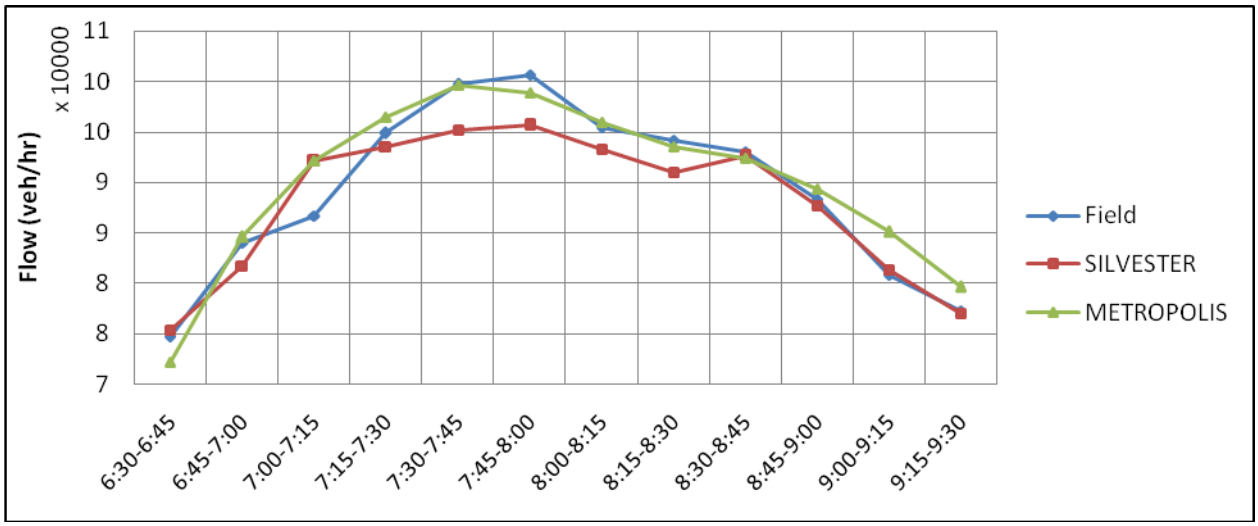


Figure 3: Distribution of total hourly flow in 59 calibration links

Figure 2 and 3 show that calibration results for both the models are good. The R^2 value suggest that the calibrated SILVESTER and METROPOLIS model can capture about 95% of the observed variability in link flows on the 59 calibration links. The observed and modelled distribution of flow by 15 minutes intervals indicates that the models are capable of predicting the temporal distribution of flow.

3.3. Validation and comparison of model results in the baseline situation

The aggregate simulation results are presented in Table 5. The term 'cordon' refers to the screen line along which the charging gates are located in the situation with congestion charging (the specification is given in Section 4.1). During calibration the demand was adjusted to have similar link flows on the selected links (shown in Figure 1). The calibration process is different for two models as described in previous section. With the same demand METROPOLIS showed much lower flow in the selected links than field observation with low congestion. To have the same flow over the selected links, the demand in METROPOLIS was increased. Hence, flow over the cordon remains close for two models but the number of car trips per hour is 19.5% larger in METROPOLIS

Both models show similar congestion percentage[100]. METROPOLIS shows lower network speed which is the reason behind higher travel time.

Table 5: Aggregate result for SILVESTER and METROPOLIS

	SILVESTER	METROPOLIS
Flow over the cordon (veh/hr)	35 611	35 651
Mean travel time (min)	19.0	20.8
Congestion (%)	41.1	41.3
Speed (km/hr)	39.3	34.9
Number of car trips starting between 6:30-9:30	280 801	335 337
Mileage (10^6veh-km)	3.49	4.08

For validation of the models the travel time in 11 selected road sections are calculated and compared with field result. Position of the road sections are shown in Figure 1. Two scatter plots for field and simulated travel times for SILVESTER and METROPOLIS model as presented in Figure 4. The validation result for METROPOLIS model is closer to the observed data than for SILVESTER model. The total travel time in these 11 sections before charging was 51.17 min as obtained from field. METROPOLIS predicted the total travel time as 53.45 min, while SILVESTER predicted 47.71 min.

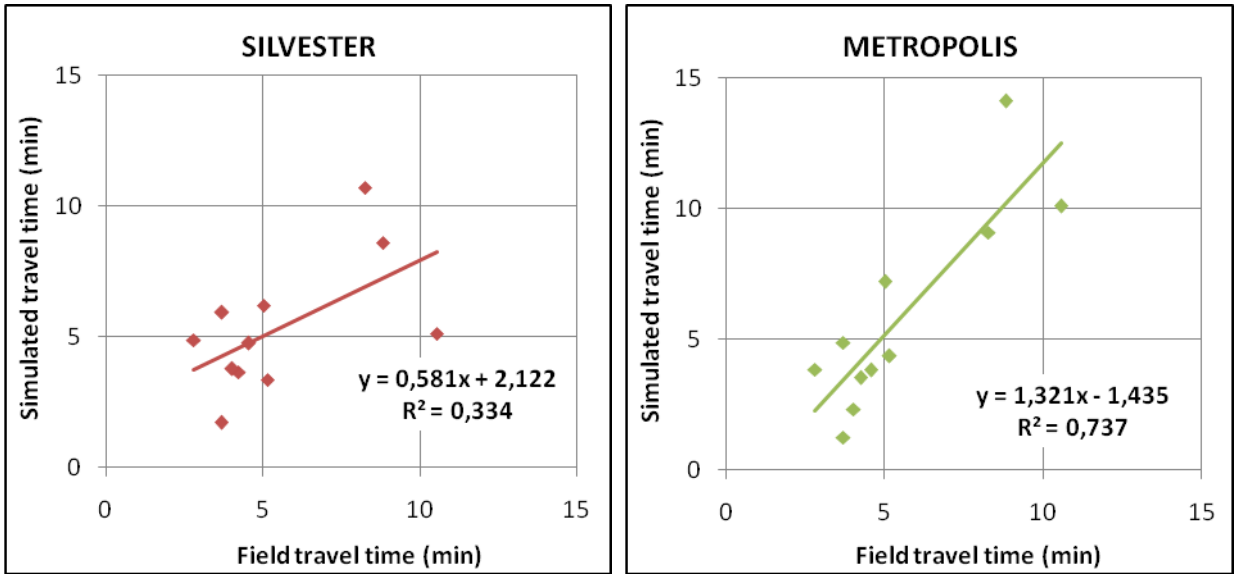


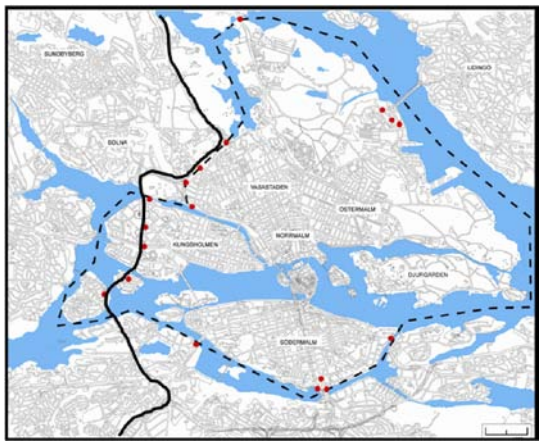
Figure 4: Field vs. Simulated travel time on 11 road sections before charging situation

[100] Congestion percentage is the relative difference between the actual total travel time and the total free-flow travel time.

4. Application to Stockholm congestion charging

4.1. Stockholm congestion charging scheme

Stockholm is the capital and the largest city of Sweden. A large fraction of the morning rush hour traffic is directed towards the central areas and is concentrated on a few main roads. A time-dependent congestion charging system has been made permanent in Stockholm from August 1, 2007 after a full scale six months trial performed in 2006. The charging system is implemented as a cordon around the city. The cordon surrounds an area with a diameter of approximately 5 km and with about 315000 people living inside. The position of the tolling stations is shown in Figure 5. The owners of all non-exempted cars driving through the cordon between 6.30 am to 6.30 pm are charged between 10 and 20 SEK depending on the time of day.



Time	Congestion charge (SEK)
06:30–06:59	10
07:00–07:29	15
07:30–08:29	20
08:30–08:59	15
09:00–15:29	10
15:30–15:59	15
16:00–17:29	20
17:30–17:59	15
18:00–18:29	10
18:30–06:29	0

Figure 5: The charging points (red dots) and the charging schedule (table to the right).

4.2. Response to congestion charging

In this section the simulated demand and system response to congestion charges are compared for the two models. The aggregate results are presented in Table 6. SILVESTER shows stronger modal shift than METROPOLIS. Field observation shows 18.1% decrease in traffic flow over the cordon. SILVESTER overestimates the flow change while METROPOLIS underestimates it. Change in other parameters like travel time, congestion and speed are very similar for the models.

Table 6: Change in aggregate results due to charging

	SILVESTER	**METROPOLIS**
Number of car	-5.0%	-2.6%
Flow over the cordon	-25.3%	-12.4%
Average travel time OD-par	-6.8%	-7.6%
Congestion	-20.7%	-22.9%
Speed	7.1%	7.6%
Mileage	-5.16%	-1.11%
Consumer surplus, MSEK	0.53	-0.61
Revenues, MSEK	0.91	1.27
Net benefit, MSEK	1.44	0.66

The change in consumer surplus shows how much the travelers gain or lose from the congestion charging system, before the revenues are returned to the population. In SILVESTER, the total surplus is calculated as logsum for each draw of the mixed logit simulation weighted by the number of travelers represented by the draw. In METROPOLIS, the surplus is computed as logsum for the binary mode choice and aggregated over all travelers. The consumer surplus and revenue values obtained

from METROPOLIS were normalized to the time period between 6:30 and 9:30 AM in order to compare them to the corresponding results from SILVESTER. This was done by applying the share of travelers that have preferred departure time in this period. The resulting revenue collection is lower in SILVESTER due to lower flow through the cordon in the charging scenario and the fact that METROPOLIS model does not take into account that some vehicles are exempted from charging while SILVESTER does (Kristoffersson, 2011).

The surplus includes the tolls paid by the drivers. According to the standard textbook analysis (Walters, 1961), the drivers paying the congestion charge are not fully compensated by shorter travel times whereby the change in consumer surplus shall be negative. However the standard analysis considers one link connecting one origin-destination (OD) pair with static volume-delay function and homogeneous travelers. The benefit of congestion charging may be higher in a road network with multiple OD-pairs (Verhoef and Small, 2004), when the drivers have different values of travel time savings (VTTS) (Ibid), or when they can adjust their departure time (Arnott et al., 1994). In METROPOLIS, all drivers with the same trip purpose (fixed, flexible or business) have the same VTTS while in SILVESTER the VTTS for each trip purpose is distributed on a long interval. Verhoef and Small (2004) showed that ignoring heterogeneity of VTTS in a system with a free parallel road leads to great underestimation of social benefits, by disregarding the efficiency gains due to separation of traffic. This may explain why the consumer surplus is higher in SILVESTER than in METROPOLIS.

Traffic flow in 59 selected points has been analyzed after the charge and it still shows good result for both models in comparison to field flow. The result is shown in Figure 6. Similarly travel time results after the charge for 11 selected road sections are compared with field travel time as shown in Figure 7. SILVESTER shows better R^2 than before while METROPOLIS remains at the same level.

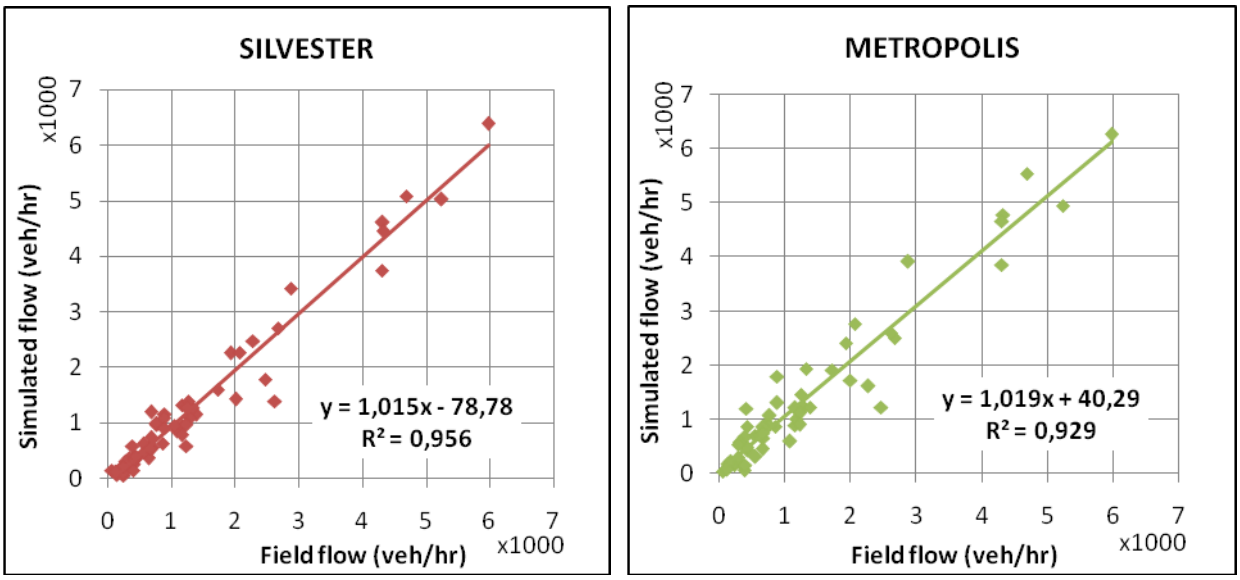


Figure 6: Field vs. Simulated flow in 59 calibration links after charging situation

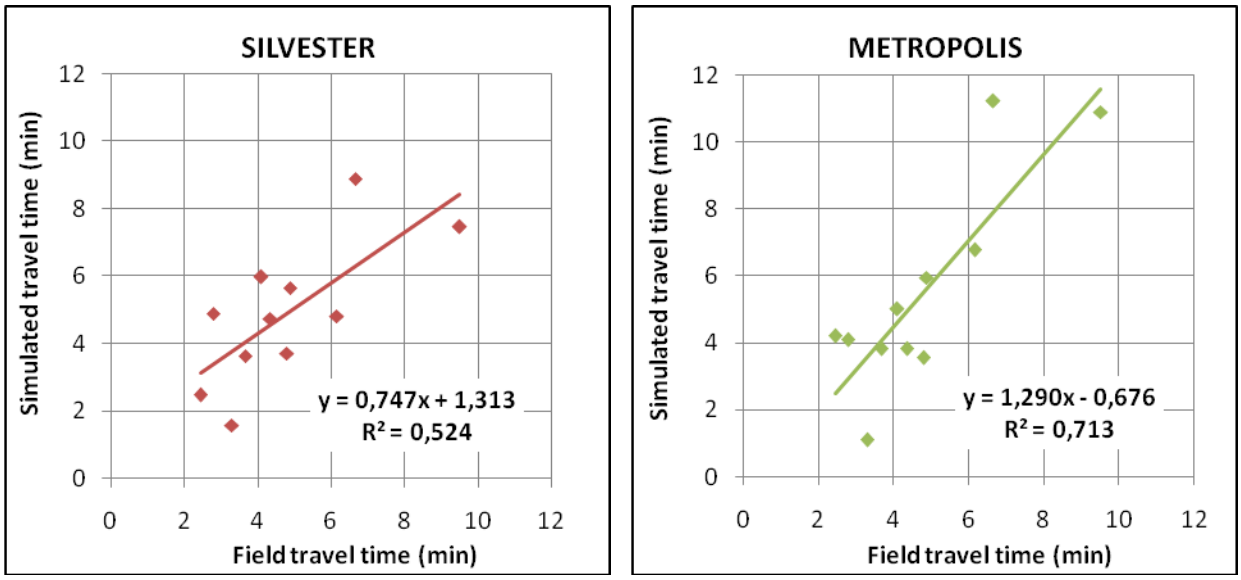


Figure 7: Field vs. Simulated travel time in 11 road sections after charging situation

In order to observe the temporal change in traffic flow over the simulation period the flow data in every 15 min interval both before and after implementation of charging are compiled. Figure 8 shows the change in total flow for 59 calibration links for each time interval. The figure shows that SILVESTER predicts higher reduction of flow during peak period than field measurement. Flow reduction in METROPOLIS is lower than field but the reduction pattern is similar to the field.

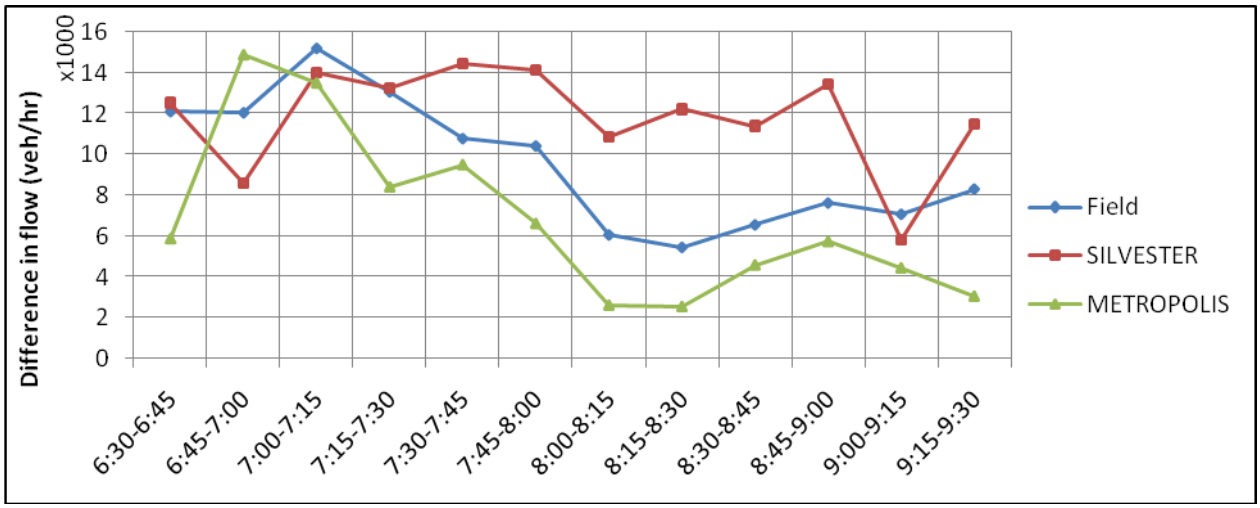


Figure 8: Temporal change in traffic flow for 59 calibration links

To observe the change in travel time due to charging in each of the 11 road sections, a linear plot is made which is presented in Figure 9. It is observed from the figure that SILVESTER model shows better prediction of travel time change than METROPOLIS model. It is worth mentioning here that the total reduction of travel time in these 11 road sections are 7.8 min as observed from field data. SILVESTER model predicted this decrease as 4.3 min whereas METROPOLIS predicted it as 2.8 min. The decrease in travel time is not so great for METROPOLIS due to two road sections: St Eriksgatan and Stora Mossen. The link St Eriksgatan is a special one. This is the only link in the city where increase of the flow was observed as a result of congestion charging. This is because an alternative route for many trips going via this link from the city would be to cross the cordon trice. So they use this link and pay just for one crossing. In spite of the flow increase the travel time actually decreased

because the conflicting flows on the intersections decreased. This is captured by CONTRAM but not by METROPOLIS and this example shows that this can be an important feature for local studies. Stora Mossen link is a continuation of St Eriksgatan and probably can be explained by the same reason.

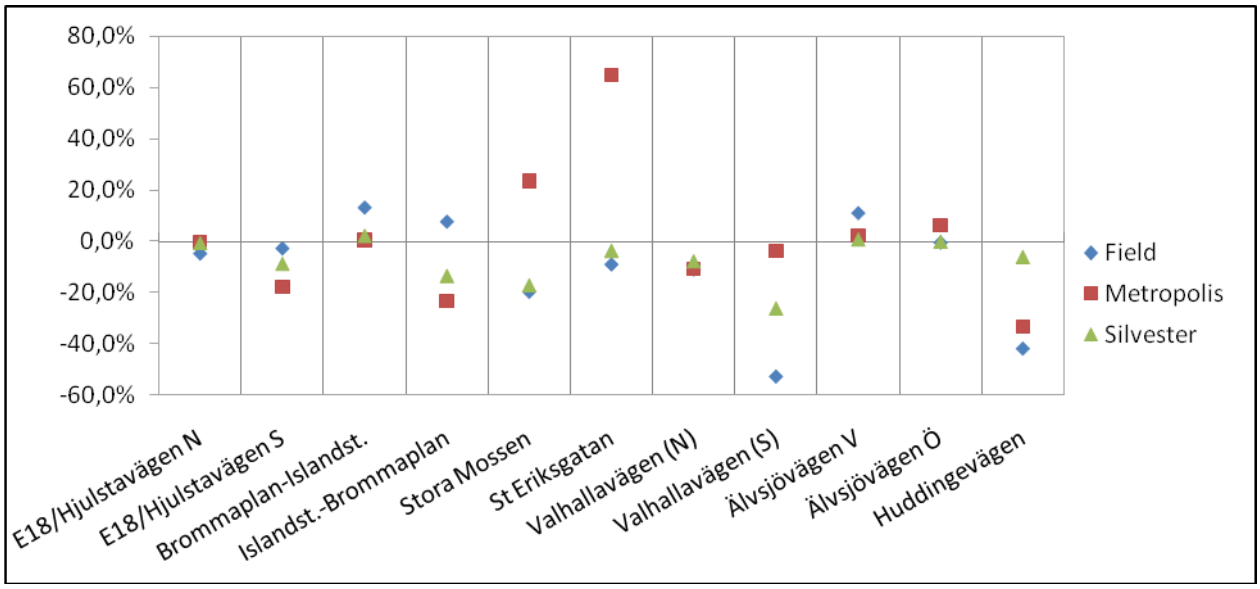


Figure 9: Change of travel time in different road sections

5. Conclusions and recommendations

Road pricing is one of the most attractive solutions to the increasingly important problem of congestion in urban areas. However, there is a strong opposition to road pricing, and therefore a need to develop reliable models to assess the different impacts of road pricing. The assessment of road pricing is usually made by a cost benefit analysis. We believe that such cost benefit analysis needs to be based on measures and indexes that are based on strong economic principles. This is the case of the two dynamic models, SILVESTER and METROPOLIS, which have been used to asses Stockholm congestion charging scheme.

The mere aggregate results are not enough for the assessment. In order to assess road pricing, one should also address its benefit (and cost) along the different dimensions (mainly congestion cost, flow, revenue, schedule delay cost, mode shift and speed). Moreover, disaggregation is needed at the user level. This is because with positive overall benefit road pricing can have a negative impact on some individuals. Often, the impacts are believed to be regressive, in the sense that poor commuters are worse off, while rich commuters (more flexible) are better off.

Validation and calibration of a dynamic models for a big city, although a large effort, is necessary in order to get a reasonably reliable assessment of the congestion charges. Not only the traffic flows on the charging locations but also on bypasses and the travel times on mayor highways and location of traffic queues have to be calibrated. After calibration of the two models for the situation without charging, we have managed to predict the impacts of Stockholm congestion charging scheme in a satisfactory manner at aggregate level. The aggregate reduction of travel times is similar between the models. However the computed reduction in flow over the cordon is rather different between the two models, one overestimating and another underestimating the flow reductions provided by field data. Note that the fit of both flow change and travel time change is very difficult to achieve in a static model. The flexibility in the dynamic model appears sufficient to fit these two fundamental measures of traffic. In this respect, we have observed a significant improvement compared to the

static model that was used for predicting the effect of congestion charges in Stockholm (Engelson and van Amelsfort, 2011).

The major response of the drivers in the two models is the shift in departure time choices due to the dynamic congestion charge. This response is clearly impossible in any static model and difficult in a dynamic assignment model. Our result indicates that the dynamic traffic models used, SILVESTER and METROPOLIS, provide satisfactory fit and predictions.

Our results provide the benefit of road pricing. Basically the benefit are negative according to METROPOLIS when the user have to pay for tolls, however, after redistribution as a lump sum, the benefits are positive. The results are more optimistic with SYLVESTER, possibly because the latter model used wider distribution of VTTS, so that the users can adjust to the changes in a more convenient and efficient manner.

Regarding differences between SILVESTER and METROPOLIS, the preliminary results indicate that the fully dynamic property of METROPOLIS with appropriate integration of scheduling and routing decisions is an advantage over the quasi-dynamic SILVESTER, since it provides flow profiles that are smoother, and therefore more in line with the smooth flow profiles of field measurements. Advantages of SILVESTER are that it has a more advanced demand model (mixed Logit compared to nested logit) and more detailed supply model (spillback and intersection interactions). This translates in the preliminary results mainly for the consumer surplus.

Acknowledgement

The authors are grateful to Fabrice Marchal for advices regarding calibration of METROPOLIS. The research was financed by Sweden, France, Denmark, Finland and Switzerland within ERA-NET TRANSPORT under the theme SURPRICE (“Road User Charging for Passenger Vehicles”).

References

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

Arnott, R., A. de Palma and R. Lindsey, (1994), The welfare effects of congestion tolls with heterogeneous commuters, *Journal of Transport Economics and Policy*, 28(2), 139–161.

Bierlaire, M., (2003), BIOGEME: a free package for the estimation of discrete choice models, In *Proceedings of the 3rd Swiss Transportation Research Conference*, Ascona, Switzerland, March 2003.

Börjesson, M., (2006), Issues in Urban Travel Demand Modelling: ICT Implications and Trip Timing Choice, Doctoral Thesis, KTH, Stockholm.

Börjesson, M., (2008), Joint RP-SP data in a mixed logit analysis of trip timing decisions, *Transportation Research Part E*, 44(6), 1025–1038.

de Palma, A. and M. Fosgerau, (2011), Dynamic and Static Congestion Models: a Review. In *Handbook in Transport Economics*, A. de Palma, R. Lindsey, E. Quinet et R. Vickerman, (eds.), Edgar Elgard, 2011.

de Palma, A. and R. Lindsey, (2006), Modelling and evaluation of road pricing in Paris, *Transport Policy*, 13(2), 115–126.

de Palma, A., P. Hansen and M. Labbé, (1990), Commuters' Paths with Penalties for Early or Late Arrival Time, *Transportation Science*, 24(4), 276–286.

de Palma, A. and F. Marchal, (2002), Real cases applications of the fully dynamic METROPOLIS tool-box: an advocacy for large-scale mesoscopic transportation systems, *Networks and Spatial Economics*, 2(4), 347–369.

de Palma, A., M. Kilani and R. Lindsey, (2005), Congestion pricing on a road network: A study using the dynamic equilibrium simulator METROPOLIS, *Transportation Research Part A*, 39(7-9), 588–611.

de Palma, A., F. Marchal and Y. Nesterov, (1997), METROPOLIS: Modular system for dynamic traffic simulation, *Transportation Research Record*, 1607, 178–184.

Eliasson, J. and L.G. Mattsson, (2006), Equity effects of congestion pricing: Quantitative methodology and a case study for Stockholm, *Transportation Research Part A*, 40(7), 602–620.

Engelson, L. and D. van Amelsfort, (2011), The role of volume-delay functions in forecast and evaluation of congestion charging schemes, application to Stockholm, In *Proceedings of the Kuhmo Nectar Conference*, Stockholm, June 2011.

Glazer, A. and E. Niskanen, 2000, Which consumers benefit from congestion tolls?, *Journal of Transport Economics and Policy*, 34(1), 43–53.

Koh, A. and S. Shepherd, (2006), Issues in the modelling of road user charging, Distillate Project F, Appendix A, ITS Leeds, Available at: http://www.its.leeds.ac.uk/projects/distillate/outputs/Deliverable%20F%20Appendix%20A.pdf [Accessed May 30, 2011].

Kristoffersson, I., (2011), Impacts of time-varying cordon pricing: Validation and application of mesoscopic model for Stockholm, *Transport Policy*, In Press, Available online 6 August 2011, DOI: 10.1016/j.tranpol.2011.06.006.

Kristoffersson, I. and L. Engelson, (2008), Estimating Preferred Departure Times of Road Users in a Real-Life Network, In *Proceedings of the European Transport Conference*, Leeuwenhorst Conference Centre, October 2008.

Kristoffersson, I. and L. Engelson, (2009), A dynamic transportation model for the Stockholm area: Implementation issues regarding departure time choice and OD-pair reduction, *Networks and Spatial Economics*, 9(4), 551–573.

Kristoffersson, I. and L. Engelson, (2011), Alternative road pricing schemes and their equity effects: Results of simulations for Stockholm, In *Proceedings of the TRB 90th Annual Meeting*, Washington, D.C., January 2011.

Nielsen, O.A., A. Daly and R. Frederiksen, (2002), A stochastic route choice model for car travellers in the Copenhagen region, *Networks and Spatial Economics*, 2(4), 327–346.

Odeck, J., J. Rekdal and T. Hamre, (2003), The socio-economic benefits of moving from cordon toll to congestion pricing: The case of Oslo, In *Proceedings of the TRB 82nd Annual Meeting*, Washington, D.C., January 2003.

Pigou, A.C., (1920), *The Economics of Welfare*, 4th. London: Macnillam.

Rich, J. and O.A. Nielsen, (2007), A socio-economic assessment of proposed road user charging schemes in Copenhagen, *Transport Policy*, 14(4), 330–345.

Small, K., (1983), The incidence of congestion tolls on urban highways, Journal of urban economics 13(1), 90–111.

Taylor, N., (2003), The CONTRAM dynamic traffic assignment model, *Networks and Spatial Economics*, 3(3), 297–322.

Verhoef, E. and K. Small, (2004), Product differentiation on roads, *Journal of transport economics and policy*, 38(1), 127-156.

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

Walters, A., (1961), The theory and measurement of private and social cost of highway congestion, *Econometrica: Journal of the Econometric Society*, 29(4), 676-699.