THEMA
théorie économique, modélisation et applications

THEMA Working Paper n°2022-06
CY Cergy Paris Université, France

Tolls vs tradable permits for managing travel on a bimodal congested network with variable capacities and demands

Robin Lindsey, André de Palma, Pouya Rezaeini

November 2022 version

Tolls vs tradable permits for managing travel on a bimodal congested network with variable capacities and demands

November 13, 2022

Robin Lindsey[a], André de Palma[b], Pouya Rezaeinia[c]

[a] Corresponding author: Sauder School of Business, University of British Columbia, 2053 Main Mall, Vancouver, BC V6T 1Z2, Canada. E-mail: robin.lindsey@sauder.ubc.ca,

[b] CY Cergy-Paris University, Thema, France. E-mail: andre.de-palma@cyu.fr

[c] Sauder School of Business, University of British Columbia, 2053 Main Mall, Vancouver, BC V6T 1Z2, Canada. E-mail: pouya.rezaeinia@sauder.ubc.ca

Abstract

Congestion pricing has long been considered an efficient tool for tackling road traffic congestion, but tolls are generally unpopular. Interest is growing in tradable permits as an alternative. Tolls and tradable permits are interchangeable if travel conditions are unchanging, but not if conditions vary and tolls and permit quantities are inflexible and cannot be adapted to current conditions. We compare the allocative efficiency of tolls and tradable permits on a bimodal network under uncertainty. Road links are congestion prone and public transit may be crowded. Road traffic entering a cordon area around the downtown is controlled using either a toll or a tradable permit. Two groups of travelers can drive or take transit. Group 1 travels downtown, and if it drives it must either pay the toll or use a permit. Group 2 travels to a suburb, and can avoid the cordon by taking a bypass. All demand and cost parameters of the model can vary, either systematically or irregularly. Each parameter combination constitutes a state. Travelers learn the state in advance, and adapt their mode and route choices accordingly. A planner minimizes expected total travel costs by either setting the level of the toll or choosing the quota of permits to distribute. Two cases are considered. In the first, the toll and quota are flexible and can be adjusted to daily travel conditions. In the second, the instruments are inflexible and must be set at the same level regardless of the state. If travelers have identical preferences, the optimal flexible toll is invariant to the numbers of travelers in each group and the capacity of the link entering the cordon. The toll is robust in the sense that inflexibility causes no welfare loss if these parameters vary. By contrast, the quota is not robust.

We derive a general rule for ranking the efficiency of a fixed toll and fixed quota. We then explore a numerical example. In most instances, the fixed toll outperforms the fixed quota by a significant margin although the quota can do better in some states. We also compare the welfare-distributional effects of tolls and permits, and find that suburban travelers fare better than downtown travelers from both forms of regulation.

Keywords: traffic congestion; transit crowding; cordon toll; tradable permits; mode choice; route choice; uncertainty; equity

JEL Codes: D62; R41; R48

1 Introduction

Traffic congestion is a burden worldwide. Inrix reports that in 2019, US drivers lost nearly 100 hours to traffic congestion on average, at a personal cost of $1,377, amounting to a loss of nearly $88 billion to the US economy.[1] Congestion imposes similar costs in the UK, France, and Germany.[2] Congestion pricing has long been considered an efficient, if not the most efficient, tool for tackling traffic congestion. Advances in tolling and IT technology have reduced the costs of imposing tolls, billing motorists, and informing them about where, when, and how much they will pay. Yet, only a few city-scale congestion pricing schemes exist: in Singapore (1975), London (2003), Stockholm (2006), Milan (2008), and Gothenburg (2013). Smaller schemes have been established in Durham (2002) and Valletta (2007), and several cities including New York City have plans to introduce charges. However, many proposed schemes have failed. Public opposition and equity concerns are the most frequently cited explanations (Jaensirisak et al., 2005; Grisolía et al., 2015; Krabbenborg et al., 2020).

Due in part to opposition to tolls, quantity controls and regulations are much more widely used than tolls. These include limits on vehicle registrations, license plate number restrictions on daily use, perimeter control of traffic, driving bans, and traffic calming. These measures are often intended to curb not only congestion, but also reduce pollution, enhance safety, and reduce noise. For example, Fageda et al. (2020) report that low emission zones have been implemented in 46 out of 130 large cities in 12 European countries.

Quantity controls are generally less efficient than tolls because they do not allocate road space to those who value it the most. Tradable Permits Schemes (TPS) are a hybrid measure that combines price and quantity controls. TPS can improve on pure quantity controls by granting driving rights, and allowing agents to trade them. TPS have been used to reduce acid rain, lead, and carbon emissions; to administer Corporate Average Fuel Economy (CAFE) standards; and to allocate taxi licenses, vehicle quotas, and airport landing slots. TPS have not yet been applied to road travel. However, advances in information technology and familiarity with it are making TPS increasingly feasible. The idea is conceptually straightforward. Drivers would need a permit to make a trip, access a road, or enter a restricted area. The amount of travel would be regulated by limiting the number of permits that are issued. If permits are distributed free of charge, individuals in aggregate would not incur an additional monetary cost to travel. Furthermore, if permits are distributed on an equal-per-capita basis, lower-income households, which tend to travel less by car, would receive more permits than they use, and would earn income from selling the excess. Vertical equity could be improved further by giving lower-income households more permits.

If travel conditions are known and unvarying, a TPS can be designed to support the same travel pattern as tolls. The two instruments are then allocatively equivalent, with the TPS having a likely advantage in terms of acceptability. However, as initially shown by Weitzman (1974) in a planned-economy context, price and quantity instruments are not equivalent in a nonstationary environment if they cannot be adjusted as conditions change. Moreover, road travel conditions do fluctuate. Travel demand varies predictably by time of day, day of week, and season. Unanticipated fluctuations in capacity and demand are also common. According to US FHWA (2020), about half of total congestion delays arises from nonrecurring events such as crashes, debris, work zones, traffic signal failures, bad weather, special events, and so on. These shocks can cause substantial delays. Closure of one lane of a three-lane highway due to an incident reduces total capacity by about half.[3] Heavy rain can reduce freeway capacity by 10-15% (Van Lint et al., 2000; Chung et al., 2006; Brilon et al., 2008), and snow storms and other severe weather conditions by more than 20%. Capacity can also drop if individual drivers change lanes or brake sharply, and cause traffic flow to break down (Brilon et al., 2008). Public transit service disruptions due to strikes and other shocks are also common in some cities, and they can affect traffic congestion

[1] https://inrix.com/press-releases/2019-traffic-scorecard-us/
[2] https://internationalfleetworld.com/the-true-cost-of-congestion/
[3] Transportation Research Board (2000, Chapter 22, Table 22-6).

if transit users switch to driving.[4] Rain, snowfall, and high temperatures encourage commuters to drive rather than take public transit or walking (Belloc et al., 2022). Finally, great shocks such as the COVID pandemic can affect travel demand for extended periods (de Palma et al., 2022).

Tradable permits and tolling schemes would still be allocatively equivalent if they could adapt to all these fluctuations. Yet, with the exception of dynamic pricing on some High Occupancy Toll (HOT) lanes in the US and Israel, real-time, state-dependent road pricing has not been adopted (Lombardi et al., 2021). Several explanations have been suggested. The infrastructure, operating, and accounting costs may be too high to justify state-dependent pricing over short time horizons. People may prefer fixed charges because adapting to frequent price changes is difficult (Bonsall et al., 2007), or because the outlays are easier to predict (Li and Hensher, 2010). People may also dislike price uncertainty per se (Lindsey, 2011), and they may dislike having to pay high tolls if they have no good alternative.

The potential for an adaptable TPS is still unknown since no schemes are operational yet. Nevertheless, transactions costs militate against frequently changing either the number of permits distributed, or the number of permit units required to use particular roads or enter certain areas. Permit quotas are likely to be distributed weekly, monthly, or quarterly rather than daily, and requirements are likely to be set systematically (e.g., by day of week or season) rather than adjusted according to current travel conditions.[5]

In this paper, we assume that tolls and tradable permits allocations are inflexible, and thus cannot be adjusted on the basis of daily travel conditions. We compare the allocative efficiency of tolls and permits in the face of shocks to demand, road capacity, and transit, assuming no administration, transactions, or traveler compliance costs are incurred for either instrument. Unlike earlier studies that compare tolls and permits under variable conditions, in our model the instruments are applied over only part of the transport network. Hence, they are second-best and cannot support a first-best optimum even under stationary conditions. We summarize our main results at the end of the literature review in the next section.

2 Literature review

This paper relates to three streams of literature: on congestion pricing, on tradable permits systems, and on network reliability and robustness.

Congestion pricing

The literature on congestion pricing is vast.[6] Most work has assumed that travel conditions are known, but congestion pricing under uncertainty has received some attention. Studies of state-dependent pricing have considered tolling single links (Emmerink et al., 1998), tolling High Occupancy Toll lanes (Dong et al., 2011; Gardner et al., 2014), and tolling road networks (Gardner et al., 2011). Vosough et al. (2020) analyze predictive tolling of congestion and air pollution based on daily weather forecasts. A few studies have considered fine-grained dynamic pricing of congestion and emissions whereby vehicles are charged incremental tolls each time they traverse a road link (Kickhöfer and Kern, 2015; Agarwal and Kickhöfer, 2018). Conversely, other studies have assumed that tolls are inflexible, and derived optimal state-independent tolls when demand and/or supply

[4] London, Paris, and New York City have suffered major public transport strikes with lasting consequences. Larcom et al. (2017) demonstrate how a strike on the London underground in 2014 induced permanent changes in behaviour. Strikes are frequent in Rome (Russo et al., 2021), and recurring in other countries. For example, Rivers et al. (2020) study 105 municipal transit strikes that occurred in Canadian cities between 1974 and 2011. Transit service can also be disrupted by mechanical failures, aging infrastructure, and natural disasters (Zhu et al., 2017).

[5] Tolls are sometimes set this way. For example, toll schedules in Singapore are adjusted quarterly, and during school holidays, to maintain target speeds on arterial roads and expressways.

[6] Literature reviews are found in de Palma and Lindsey (2011), Santos and Verhoef (2011), and Mobility Pricing Independent Commission (2018). Lehe (2019) provides a comprehensive summary of urban congestion pricing schemes.

conditions vary (e.g., Verhoef et al., 1996; de Palma and Lindsey, 1998). Thus, both flexible and inflexible pricing under uncertainty about travel conditions have been studied.

Tradable permits systems

By comparison with congestion pricing, the literature on TPS for roads is relatively new, although growing rapidly.[7] Verhoef et al. (1997) were the first to consider the use of TPS to reduce road traffic congestion and other externalities. Goddard (1997) independently proposed tradable permits as a tool to control vehicle emissions. Subsequent studies have investigated many aspects of TPS for roads, including initial permits allocation, multimodal networks, time-of-day variations in permit requirements, permit duration and banking, transactions costs, loss aversion, and welfare-distributional effects with heterogeneous travelers.

In contrast, few studies have compared the efficiency of TPS and tolls under uncertainty. The extensive environmental-economics literature on price versus quantity controls under uncertainty cannot be applied directly because, unlike with externalities such as carbon emissions, drivers collectively bear the costs of congestion.[8] Among the few studies of roads, Shirmohammadi et al. (2013) analyze a small road network with uncertain travel demand and capacity. They focus on volatility in permit prices, and assume that the regulator sells permits to prevent the price from exceeding a ceiling level. They do not evaluate the efficiency of tradable permits or compare them with tolls.

de Palma et al. (2018) explore a setting in which travelers choose between driving on one of several routes and taking public transit. Travelers' choices are determined by a mixed-logit choice model, and both demand and capacity vary stochastically. The authors solve equilibrium numerically for a large combination of parameter values, and find that a TPS outperforms tolls in a majority of instances although the average difference is not large. de Palma and Lindsey (2020) study a single congestible facility. Agents decide whether to use it conditional on the state and either the usage fee or the number of permits issued. Both fees and permit quotas are constrained to be the same across states. The authors show that the relative allocative efficiency of a TPS and a fee depends on the curvature of the cost function, the nature and magnitude of demand and cost shocks, and whether the permit quota always binds. Geng et al. (2021) explore a more elaborate model in which travelers choose between driving and taking public transit. Public transit passengers are exposed to health risks as well as crowding, and both the private and external social risks of infection are uncertain. A first-best congestion toll is imposed for driving. A planner chooses between a fixed transit fare and a fixed permit quota that limits transit ridership. The authors show that the relative efficiency and welfare effects of prices and permits depend on uncertainty about the private risks, but not the social risks.

Shirmohammadi et al. (2013), de Palma et al. (2018), de Palma and Lindsey (2020), and Geng et al. (2021) all employ static models in which the dynamics of congestion and the trip-timing preferences of travelers are disregarded. Akamatsu and Wada (2017) instead adopt the Vickrey (1969) bottleneck model in which congestion takes the form of queuing. Permits are differentiated by the time at which the bottleneck is crossed. Akamatsu and Wada show that if the regulator has full information, a TPS and a time-varying toll can both support a system optimum. If the regulator has only imperfect information about demand, a system optimum can still be realized with a TPS by issuing a number of permits commensurate with bottleneck capacity. By contrast, a fixed toll cannot support the optimum because the optimal toll schedule depends on the level of demand. In this instance, a TPS has a clear advantage over tolls.

Seshadri et al. (2022) explore another dynamic model featuring a single origin-destination (OD) pair and route, stochastic demand, and multiple time periods with different toll levels and permit requirements for each period. Similar to de Palma et al. (2018), travel demand is described by a logit mixture model. Buying and

[7] See Fan and Jiang (2013), Grant-Muller and Xu (2014), and Dogterom et al. (2017) for reviews.
[8] Czerny (2008, 2010) demonstrates this point in the context of airport congestion pricing.

selling decisions of users within each period are modeled, allowing for imperfect rationality of sellers. The authors find that permits outperform a toll when congestion is highly sensitive to usage, demand is high, and daily fluctuations in demand are large.

Network reliability and robustness

Our paper relates to the literature on network reliability and robustness because it features a simple network, and travel demand, road conditions, and transit service quality are all stochastic. Early work on network reliability focused on the threat of severed links, resulting in loss of network connectivity (Taylor and D'Este, 2007). Later research shifted to degraded networks in which (as in this paper) links continue to function, but with reduced capability to handle vehicular traffic or passenger flows. As Jenelius et al. (2005) discuss, network reliability is quantified by the *vulnerability* of links to shocks, and the consequences or *importance* of shocks (e.g., as measured by increases in generalized travel costs). Links that are both vulnerable and important are said to be *critical* links (Nicholson and Du, 1994). Networks that can accommodate shocks with only modest loss of performance are deemed to be *robust*. In the numerical part of our paper, we assess how vulnerability and importance of the links in our network affect network robustness and the relative efficiency of fixed tolls and fixed permits.

This paper

This paper builds on previous work by investigating the choice between tolls and permits for controlling access to a cordoned area akin to existing cordon pricing schemes. The model features a small network with road links in series and parallel. There are two groups of travellers. Group 1 travels into the cordon area, and Group 2 travels to a suburb. Each group can either drive or take public transit that operates on a separate right of way. Two model specifications of transit service are considered. One features crowding.[9] The other lacks crowding, but assumes that travelers differ in their preferences for taking transit. In both cases, transit is an imperfect substitute for driving. Ridership is endogenous although total travel demand is fixed.

The model differs from previous work with uncertainty in that the toll or permit requirement is imposed only on the road link that enters the cordon. The analysis is therefore second-best in the sense that even a flexible toll or TPS cannot support the system optimum. All parameters characterizing demand, free-flow costs, and capacities of each road link and transit service can be stochastic, and vary either independently or in a correlated way. As in the aforementioned studies on TPS, as well as some studies of network reliability (e.g., Dalziell and Nicholson, 2001; Chen et al., 2002; Chen et al., 2007), we assume in our central case that tolls and permit quotas are inflexible, but individuals learn travel demands and supply conditions before making their travel decisions. This assumption is plausible given the wide availability of travel information from websites such as waze.com, mobile phones, social media, and other sources. Transit users can get real-time alerts from transit websites and mobile apps. Notices about transit strikes allow travelers to make alternative plans (Job et al., 2001).[10]

Among other results, we find that in certain cases the toll is robust in the sense that the optimal flexible toll is invariant to the values of certain parameters. If so, inflexibility of the toll causes no welfare loss if these parameters vary. The quota is robust in fewer instances. When neither instrument is robust, their relative performance depends on which parameters vary. Numerical examples indicate that a fixed toll outperforms a

[9]Transit crowding is a problem in many large cities (Hörcher and Tirachini, 2021). It can increase waiting time and in-vehicle travel time, reduce travel time reliability, and cause stress while accessing transit stations, and entering, riding, and exiting transit vehicles. Although transit usage fell during the COVID-19 pandemic, it has been recovering. Moreover, health safety concerns may call for continuation of physical distancing, with potentially longer delays in boarding vehicles, lower vehicle occupancies, and greater costs of a given level of crowding.

[10]For example, strikes in Rome are announced several days ahead (Russo et al., 2021). Advance warnings about severe weather are also common, such as heavy rain in Hong Kong (Lam et al., 2008).

fixed quota in most instances. The welfare loss from inflexibility of the instruments is generally quite small although the fixed quota can result in substantial losses in particular states.

The paper is organized as follows. Section 3 describes the model and its comparative statics properties. Section 4 briefly explains how the optimal toll and permit quota are derived. Section 5 examines the properties of the flexible toll and quota, and identifies instances in which they are robust to variations in demand- or supply-side parameters. Section 6 explores the welfare effects of the two instruments on individual travelers. Section 7 analyzes the fixed toll and fixed quota, and examines their relative efficiency. Section 8 explores a numerical example, and Section 9 concludes.

3 A model of a stylized city with roads and transit service

3.1 Network components and travel choices

The bimodal city network is shown in Figure 1. There is one origin, O, which could be an inner or outer suburb. Two groups of travelers live at O and travel to distinct destinations. Group $g = d$ comprises N_d individuals who travel to downtown (D).[11] Group $g = s$ comprises N_s individuals who travel to a more distant suburban workplace (S). The two groups are hereafter called downtown travelers and suburban travelers, respectively. The numbers in each group are exogenous; they do not depend on travel costs. This is plausible for work trips when telecommuting is not an option. However, the numbers can vary stochastically and independently from day to day. Transit lines serve downtown (R_d) and the suburb (R_s). Both lines operate on separate rights of way, and do not interact either with each other or with road traffic. The road network has four arcs or links: an upper link (u), a cordon toll link (t), a lower link (l), and a bypass (b). Both the toll and the permit requirement are applied to link t, hereafter called the cordon link.[12]

Downtown travelers can either drive or take transit. If they drive, they take the upper and cordon links. They do not have a route choice. N_d^R take transit and N_d^A drive, where $N_d^R + N_d^A = N_d$. Suburban travelers who drive have a choice between a direct route and an indirect route. The direct route includes the two links to downtown, followed by the lower link to the suburb. The indirect route avoids the cordon by following the upper link and then the bypass. N_s^R suburban travelers take transit, N_s^t drive the direct route, and N_s^b use the bypass, where $N_s^R + N_s^t + N_s^b = N_s$. Under toll regulation, downtown travelers and suburban travelers who cross the cordon pay the same toll. Under permit regulation, they face the same (unit) permit requirement.

Traffic volumes on the four road links are $v_u = N_d^A + N_s^t + N_s^b$, $v_t = N_d^A + N_s^t$, $v_l = N_s^t$, and $v_b = N_s^b$. The links differ in the composition of drivers who use them. The upper link is used by all three sets of drivers, the cordon link is used by downtown travelers and suburban travelers who take the direct route, the lower link is used only by suburban travelers who take the direct route, and the bypass link is used only by suburban travelers who take the bypass route and avoid driving through downtown. As discussed later, these differences in composition are partly responsible for differences in the efficiency of the inflexible toll and inflexible permit quota.

[11] The Appendix provides a notational glossary.

[12] The network in Figure 1 could represent one corridor or pie-shaped sector of a monocentric city. The cordon would then form a closed boundary around the city centre, as in a real city.

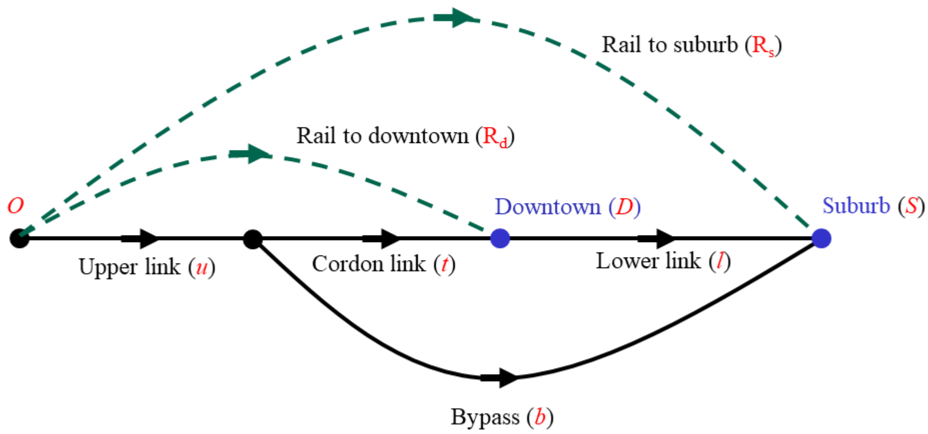

Figure 1: The network with one origin (O) and two destinations: downtown (D) and a suburb (S)

3.2 Travel costs

Driving costs

All travelers in each group incur the same cost if they drive. All link costs are linear functions of the traffic volume on them:

$$c_a\left(v_a\right) = c_{ao} + c_a' v_a,\ a = u, t, l, b,$$

where c_{ao} and c_a' are strictly positive constants. The marginal social cost of a trip on link a is $c_{ao} + 2c_a' v_a$, and the total cost is $c_a\left(v_a\right) v_a$. The c_{ao} parameters will be called *free-flow link costs*. These costs include the costs of fuel consumption, vehicle wear and tear, and any other private costs that do not depend on traffic volume. The c_a' parameters measure road-link impedances. They vary inversely with road capacity. They are exogenous, but may fluctuate from day to day. For ease of reference, we will refer to capacity shocks on the understanding that a reduction in capacity corresponds to an increase in impedance.

Let p denote the user fee for driving. Variable p is the toll with toll regulation, the equilibrium price of a permit with a TPS, and zero without regulation. The cost of driving for downtown travelers equals the cost of traversing the upper and cordon links, plus (if applicable) the toll or permit price :

$$C_d^A = c_u\left(v_u\right) + c_t\left(v_t\right) + p.$$

For suburban travelers, the cost of driving the direct route to the suburb is

$$C_s^t = c_u\left(v_u\right) + c_t\left(v_t\right) + c_l\left(v_l\right) + p,$$

and the cost of taking the bypass route is

$$C_s^b = c_u\left(v_u\right) + c_b\left(v_b\right).$$

Transit costs

Transit costs are described using two alternative specifications. In one, called *Crowded transit*, transit suffers from crowding costs. In the other, called *Diverse transit*, travelers differ in the cost they incur from taking transit and hence their preferences for transit versus driving.[13] Crowded transit differs from Diverse transit in that passengers impose a negative externality on each other. As shown below, this leads to differences in optimal modal splits for the two transit specifications as well as differences in the first-best and second-best welfare gains from intervention and the relative efficiency of tolls and quotas.

With Crowded transit, the private cost of a transit trip for each group is assumed to be a linear increasing function of ridership:

$$C_g^R\left(N_g^R\right) = C_{g0}^R + c_g^{'R} N_g^R,\ g = d, s, \tag{1}$$

where C_{g0}^R and $c_g^{'R}$ are non-negative constants that, like road capacities, are assumed to be exogenous but may fluctuate. The marginal social cost of a transit trip is $C_{g0}^R + 2c_g^{'R} N_g^R$, and the total cost is $C_g^R\left(N_g^R\right) N_g^R$.

With Diverse transit, the cost of a trip varies from person to person due to differences in accessibility to stations, physical mobility, concerns for privacy and safety, and other factors. To make the Diverse transit specification comparable with Crowded transit, it is assumed that transit costs for group g are uniformly distributed on the interval $[C_{g0}^R, C_{g0}^R + c_g^{'R} N_g]$. The individual with cost C_{g0}^R incurs the lowest cost, and the individual with cost $C_{g0}^R + c_g^{'R} N_g$ incurs the highest cost. Thus, if N_g^R individuals choose to take transit, they are the ones with costs between C_{g0}^R and $C_{g0}^R + c_g^{'R} N_g^R$. The latter individual is indifferent between taking transit and driving, and will be called the "marginal" user. Other transit users are inframarginal. Since transit users do not impose externalities on each other, the marginal social cost of a transit trip is equal to the private cost of the marginal user.

The average social costs of Crowded transit and Diverse transit can be written together as

$$\tilde{C}_g^R\left(N_g^R\right) = C_{g0}^R + f c_g^{'R} N_g^R,\ g = d, s, \tag{2}$$

where $\tilde{C}$ denotes social costs, $f = 1$ for Crowded transit, and $f = 1/2$ for Diverse transit. The downtown and suburban lines are assumed to have the same transit specification, and hence the same value of f.

Since the marginal traveler incurs the same private cost with Crowded transit and Diverse transit, the two specifications yield the same equilibria with no regulation, with the same toll, or with the same permit quota. However, for the optimal toll, the optimal quota, and the first-best optimum, optimal transit usage is higher with Diverse transit than Crowded transit since the marginal social cost of usage is lower with Diverse transit.

Crowded transit and Diverse transit are illustrated in Figure 2. With Crowded transit (shown in red), the private cost (1) for all users has a slope of $c_g^{'R}$. The marginal social cost curve (2) has twice the slope. With Diverse transit, the marginal private cost of transit (3) (i.e., the cost for the marginal user) coincides with the marginal social cost curve and has a slope of $c_g^{'R}$. The average social cost curve (4) has a slope only half as large.

[13]Transit is diverse in the sense of being different for different individuals. Idiosyncratic is an alternative term, but more cumbersome to write. Adding heterogeneity in driving preferences or values of travel time as well would complicate the analysis of route choice for suburban travelers, and further complicate the comparative statics. Given inelastic total travel demand, including preference heterogeneity for transit alone is sufficient for the purposes of the paper.

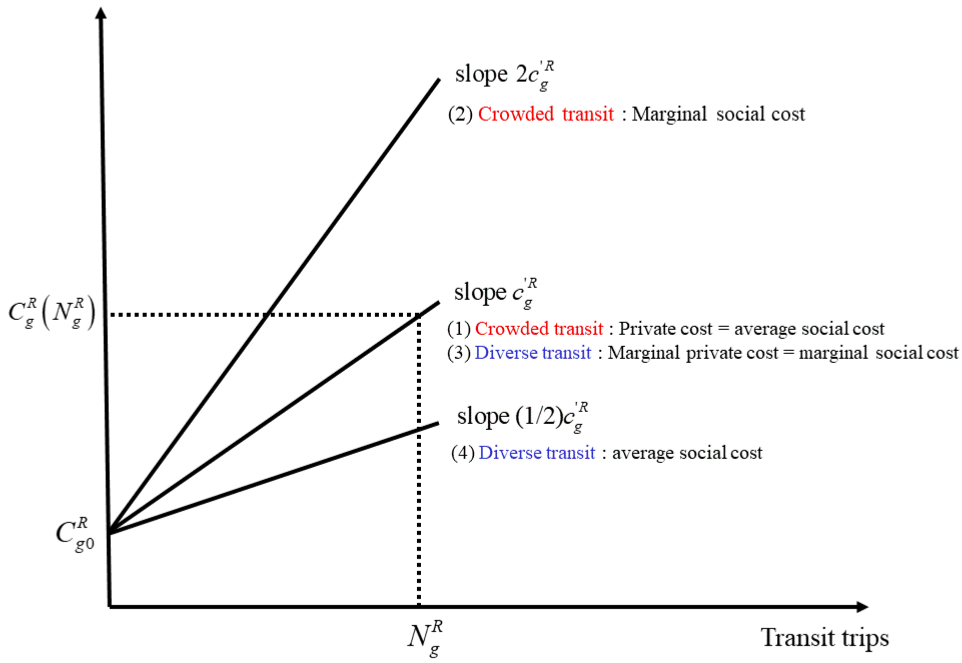

Figure 2: Crowded transit and Diverse transit

If $c_g^{\prime R} = 0$, the Crowded transit and Diverse transit specifications become the same, and the three curves in Figure 2 converge to a single horizontal line. We call this case *Uniform transit*. Since Uniform transit is more analytically tractable than Crowded or Diverse transit, and reveals some useful insights, we examine some of its analytical properties although we do not emphasize its empirical relevance.

3.3 Operation of a Tradable Permits Scheme

A Tradable Permits Scheme limits the number of vehicles that can enter the cordon area. It operates by requiring a permit to traverse the cordon link.[14] One permit is required for each trip regardless of whether the driver is bound for the downtown or the suburb. Permits are allocated free of charge, and can be traded on a competitive market with no transactions cost. In equilibrium, the permit price settles at a level such that the number of trips into the cordon area that travelers want to make matches the number of permits issued. Given no income effects in travel demand, the way in which permits are distributed does not affect equilibrium, although it does of course affect the welfare-distributional effects on individual travelers. The permit price plays the same role as a toll. If the permit quota is decreased, fewer trips can be made into the cordon area and the equilibrium permit price rises. The effect is a mirror-image of how increasing the toll reduces the number of car trips on the direct route that travelers choose to make.

3.4 Equilibrium mode and route choices

The model has 15 parameters. It is assumed that all parameters except for f (which defines the transit specification) can vary from day to day, either systematically or randomly. The parameter values realized on a given day will be called the *state*, and denoted by x. For example, if parameter c_t' takes a normal value on one day and a larger value on another day, then the two days experience different states even if all other parameter

[14] de Palma and Lindsey (2020) describe how a similar tradable permits scheme operates in a setting with no network.

values are the same. The state is assumed to be exogenous. Thus, the planner cannot increase road capacity in the event of a transit strike, or increase transit capacity during major road repairs.[15]

Individuals learn the state each day before making their travel decisions[16], and a deterministic equilibrium is assumed to be reached. Equilibrium is defined by the values of the five flows, $\mathbf{N} \equiv \left(N_d^A, N_d^R, N_s^t, N_s^b, N_s^R\right)$. Attention is restricted to interior equilibria in which all flows are strictly positive. By Wardrop's Principle, in equilibrium, modes and routes serving the same destination have the same cost for the marginal traveler. Choices with lower free-flow costs are overused because the congestion or crowding externalities on them are greater. Differences in free-flow costs therefore influence equilibrium modal and route splits, as well as optimal regulation. The free-flow costs of the various options are as follows:

$$\text{for downtown travelers : } \begin{cases} C_{d0}^R \text{ for transit,} \\ c_{u0} + c_{t0} + p \text{ for driving,} \end{cases}$$

$$\text{and for suburban travelers : } \begin{cases} C_{s0}^R \text{ for transit,} \\ c_{u0} + c_{t0} + c_{l0} + p \text{ for driving the direct route,} \\ c_{u0} + c_{b0} \text{ for taking the bypass.} \end{cases}$$

Differences in free-flow costs create the following three biases which can be positive or negative. For downtown travelers, there is a bias in favour of driving rather than taking transit:

$$B_d^R = C_{d0}^R - (c_{u0} + c_{t0} + p) \,. \tag{3}$$

For suburban travelers, there is a bias in favour of taking the bypass rather than transit:

$$B_s^R = C_{s0}^R - (c_{u0} + c_{b0}) \,, \tag{4}$$

and a bias in favour of taking the direct route rather than the bypass:

$$B_s^b = c_{b0} - (c_{t0} + c_{l0} + p) \,. \tag{5}$$

The three biases, B_d^R, B_s^R, and B_s^b, play a major role in governing equilibrium flows, as well as optimal tolls and permit quotas. Formulas for the flows can be derived using symbolic software, but except for Uniform transit they are unwieldy.

The comparative statics effects of parameter values on equilibrium flows are derived by totally differentiating the equilibrium conditions with respect to each parameter in turn. Table A.1 in the Appendix lists the signs of the effects on traveler flows and cordon link traffic volume for the equilibrium with either no regulation or an exogenous toll. Except for variations in N_d, the comparative statics signs for N_d^R are opposite to those for N_d^A. Attention is focused here on the more important and unexpected effects.

(a) The toll (τ) and the free-flow cost of cordon link (c_{t0})

The comparative statics effects of the toll and the free-flow cost of the cordon link are the same, and discussion is limited to the toll. As expected, an increase in the toll reduces driving by downtown travelers and the total number of travelers who use the cordon link. Suburban travelers reduce usage of the tolled route, too, if transit is either crowded or diverse (in which case $c_d'^R + c_s'^R > 0$), and some switch to the bypass. Less

[15]In practice, temporary capacity adjustments may be possible. Job et al. (2001) mention instances in which road capacity was increased during transit strikes by prohibiting on-street parking, or opening bus lanes to carpools.

[16]Some individuals may travel every day, others may travel regularly on certain days, and still others may vary the number of trips they make from day to day, or week to week.

obviously, suburban travelers increase their total driving if $c_d'^R c_b' - c_u' c_l' < 0$. For want of a better term, we call this the *elastic downtown transit condition*. An increase is more likely if $c_d'^R$ is small (i.e., with Crowded transit downtown transit is not prone to crowding, or with Diverse transit preferences are not highly differentiated), because downtown transit demand is then elastic and travelers reduce driving a lot if the cost of driving rises. An increase is also more likely if the bypass is not prone to congestion (c_b' is small) because suburban travelers can increase their usage of the bypass without a large increase in cost. Finally, an increase is more likely if the upper and lower links are congestion-prone (i.e., c_u' and c_l' are large) because a given reduction in downtown traveler traffic on the direct route decreases the cost of using it significantly.[17]

If a higher toll induces more suburban travelers to drive, they actually gain from a toll increase despite the fact that some of them use the direct route and pay the toll. This result is driven by network effects that are absent in earlier work featuring models with no network and/or single groups of travelers.[18] Table A.1 also shows that if the elastic downtown transit condition holds, usage of the direct route unexpectedly declines with both the number of suburban travelers (N_s) and the free-flow cost of suburban transit (C_{s0}^R).

(b) Number of downtown travelers (N_d)

An increase in the number of downtown travelers induces more of them to take transit. If downtown transit is crowded or diverse (i.e., $c_d'^R > 0$), downtown driving also increases. Suburban travelers respond by reducing usage of the direct route and increasing usage of transit. Less predictably, usage of the bypass falls if $c_s'^R c_t' - c_u' c_l' < 0$: a condition similar to the elastic downtown transit condition which will be called the *elastic suburban transit condition*. When it holds, suburban travelers prefer to shift to transit rather than take the bypass which requires use of the upstream link that has become more congested with downtown drivers.

(c) Free-flow cost of the lower link (c_{l0})

An increase in free-flow cost of the lower link discourages suburban travelers from taking the tolled route. This reduces the cost of driving for downtown travelers, who do not use the lower link. Consequently, they drive more and benefit from the increase in lower-link cost.

(d) Free-flow cost of the bypass (c_{b0})

An increase in free-flow cost of the bypass reduces use of the bypass, and increases suburban travelers' use of their other two options. If the elastic suburban transit condition is satisfied, downtown travelers drive more and end up better off. They benefit from the reduction of suburban traffic taking the bypass that uses the upper link, but lose from the increase in suburban traffic on the direct route that uses the upper and cordon links. The former effect dominates if the elastic suburban transit condition holds.

The last two comparative statics results show that downtown travelers can benefit from a deterioration in road travel conditions. In both cases, they do not use the link that is adversely affected. However, the cases differ in that the lower link runs in series with the direct route that downtown travelers use, whereas the bypass route overlaps their route on the upper link, and runs in parallel downstream the rest of the way. This illustrates the diversity of effects that materialize on the network.

[17] The result is easier to see if downtown transit is uniform (i.e., $c_d'^R = 0$). In that case, the cost of driving cannot change, and the combined cost of using the upper and cordon links net of toll must fall by the amount of the toll increase. Since all traffic on the cordon link is fed by the upper link, flow on the upper link must fall. This makes the bypass cheaper, so that suburban travelers use it more, and reduce usage of their two alternatives.

[18] It is well known that drivers with high values of time can gain from a toll (see, for example, Layard 1977). However, all travelers in the model here incur the same costs of driving.

4 Regulation to control congestion and crowding

Given traffic congestion, transit crowding (in the case of Crowded transit), and biases in mode and route choices, the total costs of travel are not, in general, minimized in the unregulated equilibrium. Total costs, TC, equal the combined travel costs incurred by downtown and suburban travelers, or, equivalently, the combined costs of driving and using transit:

$$TC = \sum_a c_a\left(v_a\right) v_a + \sum_g \tilde{C}_g^R\left(N_g^R\right) N_g^R. \tag{6}$$

The first-best optimum for a given state can be derived by minimizing total costs in Eq. (6) subject to the applicable constraints. The comparative statics properties of the first-best optimum are similar to those for the optimal toll, and to save space are not reported. The first-best optimum entails an optimal modal split for downtown travelers, and optimal modal and route splits for suburban travelers. These three conditions cannot be satisfied by only controlling traffic flow on the cordon link. Hence, the optimal toll and optimal quota are only second-best optimal. To derive them, it is useful to write total costs in terms of the flows, $\mathbf{N}$, and the state, x.

With toll regulation, the flows depend on the level of the toll, τ, and the state. Accordingly, flows will be written as $\mathbf{N}\left(\tau, x\right)$, and total costs as

$$TC\left(\mathbf{N}\left(\tau, x\right), x\right). \tag{7}$$

With permit regulation, flows depend on the quota, Q, and x. The quota affects flows through the equilibrium price: $q\left(Q, x\right)$. Total costs can thus be written

$$TC\left(\mathbf{N}\left(q\left(Q, x\right), x\right), x\right). \tag{8}$$

In Section 5 we assume that the toll and quota are flexible, and can be adapted to the state. We refer to the optimal toll as the *flexible toll*, and the optimal quota as the *flexible quota*. We examine how the flexible toll and flexible quota vary with parameter values. If an instrument is independent of a parameter, it is robust to variations in the value of the parameter, and inflexibility creates no efficiency loss. Identifying such instances helps to rank the efficiency of the two instruments when they cannot be adapted to the state.

5 The flexible toll and flexible quota

5.1 The flexible toll

General properties

Let $\tau^o\left(x\right)$ denote the flexible toll, where superscript o denotes optimal. The flexible toll is chosen to minimize total costs in Eq. (7):

$$\tau^o\left(x\right) = \underset{\tau}{Arg\min}\, TC\left(\mathbf{N}\left(\tau, x\right), x\right). \tag{9}$$

The analytics of the toll resemble second-best pricing of two parallel routes in Verhoef et al. (1996), but they are more complex and the formula for the flexible toll is very complicated for the general case with parameter f. With possible exceptions, an increase in the toll reduces congestion on the upper, cordon, and lower links, and increases congestion on the bypass and crowding on the transit lines. The toll is chosen to balance these externalities, each of which can be alleviated or exacerbated when a parameter changes. The derivatives of the toll are characterized in the Appendix. The signs of the derivatives are listed in Table 1. For ease of comparison, the corresponding properties of the flexible quota, considered in Section 5.2, are listed in juxtaposition.

Uniform transit: With Uniform transit, the flexible toll is

$$\tau^o = \frac{B_d^R}{2} - \frac{c_u' B_s^R}{2(c_u' + c_b')} = \frac{C_{d0}^R - c_{u0} - c_{t0}}{2} - \frac{c_u'(C_{s0}^R - c_{u0} - c_{b0})}{2(c_u' + c_b')}, \tag{10}$$

where B_d^R in Eq. (3) is evaluated with $p = 0$. The first term on the RHS of Eq. (10) corrects for the bias of downtown travelers toward driving rather than taking transit. Since this bias is a function of fixed costs, it does not constitute an external cost per se but it leads to excessive driving by downtown travelers on the direct route. The second term in Eq. (10) corrects for the bias of suburban travelers toward taking the bypass rather than transit, given in Eq. (4). Again, this bias does not constitute an external cost but it leads to excessive driving by suburban travelers on the bypass route. Note that if both the biases were zero, the toll would be zero because, with Uniform transit, everyone in both groups would take transit and the external costs of congestion on the direct and bypass routes would be zero.[19]

Table 1: Comparative statics effects of parameters on the flexible toll and flexible quota

Parameter	Uniform transit		Crowded transit		Diverse transit	
	Toll	Quota	Toll	Quota	Toll	Quota
N_d	0	0	0	+	+	+
N_s	0	0	0	$Sgn[c_d'^R c_b' - c_u' c_l']$	+	
c_{u0}	−	−	−	−	−	−
c_{t0}	−	−	−	−		−
c_{l0}	0	0	−	−	−	−
c_{b0}	+	+	+			
c_u'	−	−	$Sgn\left[c_s'^R\left(B_s^b - B_d^R\right) - \left(c_d'^R + c_l'\right)B_s^R\right]$			
c_t'	0	−	0			
c_l'	0	0	$Sgn[-\left(c_u' + c_b'\right)\left(B_s^R + B_s^b - B_d^R\right)$ $+c_s'^R\left(B_d^R - B_s^b\right)]$			
c_b'	+	+	$Sgn\left[c_s'^R\left(B_d^R - B_s^b\right) + \left(c_d'^R + c_l'\right)B_s^R\right]$			
C_{d0}^R	+	+	+	+	+	+
C_{s0}^R	−	−	$Sgn[c_d'^R c_b' - c_u' c_l']$	$Sgn[c_d'^R c_b' - c_u' c_l']$		
$c_d'^R$	N/A	N/A	$Sgn[\left(c_u' + c_b'\right)\left(B_s^R + B_s^b - B_d^R\right)$ $+c_s'^R\left(B_s^b - B_d^R\right)]$			
$c_s'^R$	N/A	N/A	$Sgn[c_d'^R c_b' - c_u' c_l']$			

Cells with complicated formulas are left blank.

With Uniform transit, the flexible toll in Eq. (10) is independent of the numbers of travelers (N_d and N_s),

[19] If the second bias in Eq. (10) is stronger than the first, the toll is negative. A negative toll seems unlikely in practice, and no such instances were encountered in the numerical examples described later.

the free-flow cost on the lower link (c_{l0}), and capacities of the tolled and lower links (c'_t and c'_l). Hence, the toll is robust to fluctuations in any combination of these five parameters. Explanations for some of the less intuitive effects are given in the Appendix.

Crowded transit: Unlike with Uniform transit, with Crowded transit the toll depends on the two lower-link parameters (c_{l0} and c'_l). Nevertheless, the toll is still independent of the numbers of travelers and cordon-link capacity, c'_t. This is because these parameters do not affect the biases that determine the optimal toll.[20]

Diverse transit: With Diverse transit, all derivatives are non-zero in general. Formulas for most of the signs are too complicated to interpret easily, and most of the cells in Table 1 are thus left blank.

5.2 The flexible quota

Similar to the flexible toll, the flexible quota is chosen to minimize total costs conditional on the state, given in Eq. (8). The flexible quota, $Q^o(x)$, solves

$$Q^o(x) = \underset{Q}{Arg\min}\; TC\left(\mathbf{N}\left(q\left(Q, x\right), x\right), x\right). \tag{11}$$

With Uniform transit, the signs for the toll and quota in Table 1 are identical except for cordon-link capacity, c'_t. An increase in c'_t does not change the flexible toll because it does not affect the biases that determine the optimal toll. In contrast, the flexible quota decreases because the direct route becomes less attractive to both downtown and suburban travelers. Consequently, optimal traffic volume on the cordon link declines. Thus, the toll is robust whereas the quota is not.

With Crowded transit, the quota is not robust to any parameters. Unlike for the toll, it increases with the number of downtown travelers, N_d. It decreases with the number of suburban travelers if the elastic downtown transit condition, $c'^R_d c'_b - c'_u c'_l < 0$, holds. In this case, letting additional suburban travelers take the direct route is undesirable if the upper and lower links are congestion-prone (i.e., c'_u and c'_l are large), and downtown transit is capacious so that downtown travelers can switch to transit at a relatively low cost. Table 1 reveals that the sign of $c'^R_d c'_b - c'_u c'_l$ also determines how the fixed cost of transit, C^R_{s0}, affects the toll and the quota, and how suburban transit capacity, c'^R_s, affects the toll. With Diverse transit, most of the signs that can be established are the same as for Crowded transit.

6 Welfare effects of regulation on individual travelers

Attention so far has focused on the aggregate effects of the toll and TPS. From both a public policy and a political economy standpoint, it is also important to examine the impacts on individuals (e.g., De Borger and Russo, 2018). Three assumptions will be made to limit the complexity of analysis. First, all travelers receive the same permit allocation regardless of their destination and choice of travel mode.[21] Since each traveler makes only one trip per day, and some take transit rather than drive, each receives only a fraction of a permit. Such a system could be implemented by distributing a monthly quota of permits to all residents of a given area; perhaps conditional on their having to commute to work. Second, while permits are given to travelers free of charge, travelers do not benefit from the revenues generated by tolling. This assumption gives a TPS an advantage over a toll as far as acceptability.[22] Finally, variability in travel conditions is ignored. The equity of

[20] Formulas for the flexible toll and flexible quota with Crowded transit are derived in the Appendix.

[21] Allocations could be differentiated for equity reasons. For example, Liao et al. (2022) consider giving more permits to drivers with lower values of travel time.

[22] Equivalence could be restored by assuming that toll revenues are returned to all travelers as an equal per-capita lump sum.

the two instruments can then be compared without the complicating influence of uncertainty which is unlikely to alter the qualitative conclusions.[23]

The results of the welfare comparison depend on the type of transit service. With either Uniform or Crowded transit, there are only two discrete types of travelers: downtown and suburban. With Diverse transit, travelers also differ in their individual costs of taking transit. The three types of transit are considered in turn, and the results are summarized in Table 2.

Uniform transit: With Uniform transit, no traveler is affected by a toll because equilibrium private travel costs are determined by the fixed costs of transit. If the toll is raised, travelers switch from driving to transit until the cost of driving drops back to its original level. Hence, the toll can be increased progressively from 0 up to its chosen level without affecting private costs. With a TPS, private travel costs inclusive of the cost of obtaining a permit (or the opportunity cost of not selling it) are also unaffected. However, travelers who choose to use transit can sell their permits at the equilibrium price. Hence, all travelers end up better off with a TPS.

Crowded transit: As shown in Table A.1, downtown transit ridership increases with the toll. Since the cost of Crowded transit increases with ridership, and everyone has the same preferences, tolling makes downtown travelers worse off. Table A.1 also shows that a toll reduces use of suburban transit if the elastic downtown transit condition holds, and increases it otherwise. In the former case, tolling makes suburban travelers better off because their trip cost drops.

With a TPS, travelers in aggregate do not incur an out-of-pocket cost. Since the road network is used more efficiently, travelers gain in aggregate. It is possible, but tedious, to show analytically that suburban travelers always gain. However, as shown numerically in Section 8.2, downtown travelers can lose.

Diverse transit: Assessing welfare effects for Diverse transit is more complicated than for Crowded transit because transit preferences are heterogeneous. Travelers in each group can be divided into segments according to their mode choice before and after regulation. There are four possibilities: (*Transit*, *Transit*), (*Transit*, *Drive*), (*Drive*, *Transit*), and (*Drive*, *Drive*). Segment (*Transit*, *Drive*) does not exist for downtown travelers since regulation makes driving less attractive. Segment (*Transit*, *Drive*) is possible for suburban travelers if the elastic downtown transit condition holds. However, it does not hold in the numerical example and for brevity it is omitted. The welfare effects of regulation on the three remaining segments are as follows:

(*Transit*, *Transit*): This segment is unaffected by a toll. It is better off with a TPS because the cost of taking transit does not change, and members earn money from selling their permits.

(*Drive*, *Drive*): With tolling, the welfare effects for this segment are the same as for Crowded transit. Downtown travelers are worse off, and suburban travelers are worse off, too, unless the elastic downtown transit condition holds. With a TPS, individuals of either group in this segment could be better or worse off.

(*Drive*, *Transit*): With tolling, individuals from each group in this segment are worse off since they could have taken transit before the toll is imposed, but chose not to. The marginal individual with the lowest transit cost is indifferent between driving and transit without the toll, and neither gains nor loses from tolling. The marginal individual with the highest transit cost is indifferent between driving and transit with the toll. Tolling leaves them strictly worse off. Given the uniform distribution of Diverse transit costs in the model, the average welfare change for the (*Drive*, *Transit*) segment is half the change in welfare for the (*Drive*, *Drive*) segment. Similarly, with a TPS, the average welfare change for this segment is the average of the welfare change for (*Transit*, *Transit*) and (*Drive*, *Drive*).

[23] In the numerical example the welfare gains from intervention vary little with the type and magnitude of fluctuations in travel conditions.

Table 2: Effects of toll and TPS on individual traveler welfare

	Toll		TPS	
Type of transit	Downtown	Suburban	Downtown	Suburban
Uniform	0	0	Gain	Gain
Crowded	Lose	Lose if $c_d'^R c_b' - c_u' c_l' > 0$	?	Gain
Diverse (Segment):				
(*Transit*, *Transit*)	0	0	Gain	Gain
(*Drive*, *Drive*)	Lose	Lose if $c_d'^R c_b' - c_u' c_l' > 0$	?	?
(*Drive*, *Transit*)	Lose	Lose	Mean of (*Transit*, *Transit*) & (*Drive*, *Drive*)	.

A question mark indicates that, depending on parameter values, the segment could gain or lose.

Two general results can be gleaned from Table 2. First, suburban travelers gain more (or lose less) from regulation than downtown travelers. Second, all travelers are at least as well off with a TPS as with a toll. This suggests that a TPS would be more politically palatable than a toll, particularly among suburban travelers who can continue to drive without using the regulated part of the road network.

7 The inflexible toll and inflexible quota

We now assume that the toll and quota are inflexible, and chosen to minimize the expected value of total social costs. For brevity, the optimized toll will be called the *fixed toll*, and the optimized quota the *fixed quota*. The fixed toll, τ^*, solves:

$$\tau^* = Arg \min_{\tau} E\left\{TC\left(\mathbf{N}\left(\tau, x\right), x\right)\right\},$$

where $E\{\cdot\}$ is the expectations operator. Expectations are taken over the cumulative distribution function (CDF) of states, $F(x)$, which may be discrete or continuous. The state, x, is assumed to be bounded, and all states are assumed to yield interior equilibria. The minimum achievable expected cost with the fixed toll will be written

$$E\left\{TC\left(\tau^*, x\right)\right\}. \tag{12}$$

The fixed quota, Q^*, solves:

$$Q^* = Arg \min_{Q} E\left\{TC\left(\mathbf{N}\left(q\left(Q, x\right), x\right), x\right)\right\}.$$

The minimum achievable expected cost with the fixed quota will be written

$$E\left\{TC\left(q\left(Q^*, x\right), x\right)\right\}. \tag{13}$$

With the fixed quota in place, the equilibrium permit price, $q\left(Q^*, x\right)$, is such that the number of trips on the cordon link in state x with $p = q\left(Q^*, x\right)$ equals Q^*.

7.1 Relative efficiency of the inflexible toll and inflexible quota

The fixed toll outperforms (i.e., is more efficient than) the fixed quota if $E\left\{TC\left(\tau^*, x\right)\right\} < E\left\{TC\left(q\left(Q^*, x\right), x\right)\right\}$. The fixed quota is superior if the inequality is reversed. In general, the ranking depends on the probability distribution function (pdf) of states. Nevertheless, a ranking can sometimes be established when the pdf has a particular support by comparing the efficiency losses of the two instruments when they deviate from their respective optimal values. Let X be the (nondegenerate) support of the pdf. Suppose the fixed toll is set such that it is optimal for some state, x', which may or may not be in X. If state x is realized instead, the welfare loss from having the wrong toll is

$$L_{\tau}(x, x') = TC(\tau^{o}(x'), x) - TC(\tau^{o}(x), x). \tag{14}$$

Similarly, suppose the fixed quota is optimal for state x', but state x is realized instead. The welfare loss from having the wrong quota is

$$L_{Q}(x, x') = TC(q(Q^{o}(x'), x), x) - TC(q(Q^{o}(x), x), x). \tag{15}$$

A toll and quota are equally efficient if optimized for a state that is realized because they take the same values as their flexible counterparts. Put another way, the situation is the same as a stationary environment for which price and quantity controls are equally effective. Hence $q(Q^{o}(x), x) = \tau^{o}(x)$, and it follows that:

$$TC(q(Q^{o}(x), x), x) = TC(\tau^{o}(x), x).$$

The efficiency of the fixed toll and fixed quota can be compared using Eqs. (14) and (15). The following theorem establishes a link between their respective loss functions and their relative performance.

Theorem 1 *A fixed toll strictly outperforms a fixed quota given any CDF $F(\cdot)$ on X if $TC(\tau^{o}(x'), x) < TC(q(Q^{o}(x'), x), x)$ for all $x \in X$ with $x \neq x'$. If the inequality is reversed, a fixed quota strictly outperforms a fixed toll.*

The weak version of Theorem 1 is:

Theorem 2 *A fixed toll weakly outperforms a fixed quota given any CDF $F(\cdot)$ on X if $TC(\tau^{o}(x'), x) \leq TC(q(Q^{o}(x'), x), x)$ for all $x \in X$. If the inequality is reversed, a fixed quota weakly outperforms a fixed toll.*

Theorem 1 is proved in the Appendix. The proof of Theorem 2 is analogous.

Theorems 1 and 2 rank the two instruments by comparing their robustness to deviations from the states in which they perform optimally. Suppose, for example, that the toll and permit quantity are optimized for "ideal" conditions in which all road links and transit services operate at their design capacities. If ideal conditions occur, the two instruments support the same outcome. If, however, capacity of the upper link is degraded, both instruments will operate imperfectly. If the loss with the fixed quota exceeds the loss with the fixed toll for all possible degraded levels of upper-link capacity, then the fixed toll outperforms the fixed quota given any pdf of upper-link capacity.

By deriving the loss functions, it is possible to rank the fixed toll and fixed quota for variations in individual parameter values, as shown in Table 3. Clear-cut results exist for Uniform transit. The toll and quota are both robust w.r.t. variations in the four parameters identified in Table 1 that do not affect the flexible toll or flexible quota: N_d, N_s, c_{l0}, and c_l'. The fixed toll is also robust w.r.t. fluctuations in c_t', and it outperforms the fixed quota for fluctuations in c_u' and c_b'. The two instruments perform imperfectly but equally well for variations in all four free-flow costs as well as the fixed costs of downtown and suburban transit. With Crowded transit, the fixed toll is robust and outperforms the fixed quota for the three parameters identified in Table 1 that do not affect the flexible toll: N_d, N_s, and c_t'. The two instruments again perform equally well for variations in the four free-flow costs. With Diverse transit, tractable analytical results can be derived only for free-flow costs.

The ranking varies by link. The fixed toll is superior for fluctuations in the free-flow cost of the cordon link, the fixed quota appears to be superior for fluctuations in the free-flow cost of the lower link, and the ranking can go either way for the upper link and bypass.

Table 3: Relative efficiency of a fixed toll and a fixed quota

Parameter	Uniform transit	Crowded transit	Diverse transit
N_d	Both robust	Toll robust	
N_s	Both robust	Toll robust	
c_{u0}	Equivalent	Equivalent	Toll superior if suburban transit uniform. Quota superior if downtown transit uniform.
c_{t0}	Equivalent	Equivalent	Toll superior
c_{l0}	Both robust	Equivalent	Quota superior if one transit system uniform. Unclear if neither uniform.
c_{b0}	Equivalent	Equivalent	Toll superior if one transit system uniform. Quota can be superior otherwise.
c'_u	Toll superior		
c'_t	Toll robust	Toll robust	
c'_l	Both robust		
c'_b	Toll superior		
C^R_{d0}	Equivalent		
C^R_{s0}	Equivalent		
c'^R_d	N/A		
c'^R_s	N/A		

"Robust" means independent of parameter values. "Equivalent" means not robust but equally efficient in every state. "Superior" means more efficient in every state. Cells for which rankings cannot be established are left blank.

Table 3 provides a complete efficiency ranking of the fixed toll and fixed quota for Uniform transit. However, many of the cells are blank for Crowded and Diverse transit, including all but one of the cells for road capacity which is arguably the most important source of fluctuations on the supply side. Overall, the efficiency ranking is not very distinct. This contrasts with the linear version of the simpler model in de Palma and Lindsey (2020) in which there is only one homogeneous group of individuals, only one road link or other congestible facility, and a choice between using the facility and making an alternative choice with no external costs. In their model, the fixed toll outperforms the fixed quota in most cases. The difference between the two studies can mainly be attributed to two differences in their models. First, the model here has two groups of travelers, two transit systems, and four road links. In total, travelers have five travel options (two transit, and three driving routes), and the marginal external cost of a trip generally differs for each option. Furthermore, changes in the choices made by one group cause shifts in the inverse demand curve of the other group. Second, unlike in the de Palma and Lindsey (2020) model, neither the flexible toll nor the flexible quota can support a first-best optimum. Moreover, the instruments do not discriminate between the two traveler groups despite the fact that suburban travelers travel further, and tend to generate higher external costs. A change in a parameter value can shift not only the inverse demand curve of one or both groups, but also the composition of traffic on the network so that it affects the marginal external cost of trips.

7.2 Sensitivity of expected costs to the probability distribution of states

The total travel costs realized in a given state depend on the value of the fixed toll or fixed quota. Expected total costs also depend, of course, on the pdf of states. It is possible to show that expected total costs given the optimal fixed choice of either instrument exhibit diminishing sensitivity to the pdf in a sense that we now make precise. Let y denote the value of the toll or quota, and $TC(x,y)$ total costs in state x given y. Let $y^*(F)$ denote the optimal fixed value of y given CDF $F(\cdot)$ over any set of states:

$$y^*(F) = Arg\min_{y} E_F\left\{TC\left(x,y\right)\right\}.$$

Minimized expected costs given CDF $F(\cdot)$ are

$$ETC^*(F) = E_F\left\{TC\left(x,y^*(F)\right)\right\}.$$

Let F_1 and F_2 be two distinct probability distributions. Take any $\lambda \in (0,1)$, and define $\hat{F} \equiv \lambda F_1 + (1-\lambda)F_2$.

It then follows (see the appendix) that

$$ETC^*(\hat{F}) \geq \lambda ETC^*(F_1) + (1-\lambda)ETC^*(F_2).$$

This result is formalized in:

Theorem 3 *Expected total costs given an optimal fixed choice of the toll or quota are a concave function of the distribution of states.*

Following a similar procedure (see the Appendix), it is possible to prove a similar property for the expected welfare loss due to inflexibility of the toll or quota:

Theorem 4 *The expected welfare loss from inflexibility of the toll or quota is a concave function of the distribution of states.*

Theorems 3 and 4 establish that the welfare loss from adverse shocks increases less than in proportion to their probabilities.[24] For example, suppose the probability of a 50% capacity reduction of the upper link rises from 0.1 to 0.2, while the probabilities of all other types of shocks remain unchanged. Then the increase in both expected total costs and the welfare loss due to inflexibility of the toll or quota less than doubles. We comment on this result in the following section featuring numerical examples. The underlying reason for decreasing sensitivity with respect to the probabilities of shocks is that the planner can adapt by adjusting an instrument appropriately. For instance, if the lower link becomes more susceptible to disruptions, the cordon toll can be raised or the quota reduced so that fewer suburban travelers take the direct route and suffer from a disruption.

[24] The same is not true of a shock itself: the expected loss can be a concave or convex function of the magnitude of the shock.

8 Numerical examples

8.1 Setup

Although a number of analytical results have now been derived, gaps remain and the quantitative difference in performance of the toll and TPS remains unexplored. To proceed further, numerical examples are investigated. Base-case parameter values are chosen to correspond approximately with the network used in de Palma et al. (2005) to study dynamic cordon and area-based tolls on a bimodal network with traffic congestion, but no transit crowding. Values of travel time by road and transit are updated to reflect contemporary values. Other parameters are calibrated to yield relatively high congestion and crowding externalities, as well as interior equilibria for all regulatory regimes. The parameter values are listed in Table 4. The Appendix describes the reasoning underlying them. In brief, there are $10,000$ travelers: a number broadly representative of a corridor or pie-shaped sector of a medium-sized city. The four road links have similar capacities, but the bypass is much longer than the other three links. The transit line to the suburb is longer than the downtown line, and consequently has a higher fixed cost and impedance because trips take longer.[25]

Table 4: Base-case parameter values for stationary conditions

Parameter	Description	Value
N_d	Number of downtown travelers	$4,000$
N_s	Number of suburban travelers	$6,000$
c_{u0}	Free-flow cost of upper link (6 km)	\$2.10
c_{t0}	Free-flow cost of cordon link (6 km)	\$2.34
c_{l0}	Free-flow cost of lower link (8 km)	\$3.12
c_{b0}	Free-flow cost of bypass (21 km)	\$8.19
c'_u	Impedance of upper link	\$1.2 10^{-3} /trip
c'_t	Impedance of cordon link	\$1.2 10^{-3} /trip
c'_l	Impedance of lower link	\$1.2 10^{-3} /trip
c'_b	Impedance of bypass	\$1.6 10^{-3} /trip
C^R_{d0}	Fixed cost of using downtown transit	\$13.40
C^R_{s0}	Fixed cost of using suburban transit	\$17.00
c'^R_d	Impedance of downtown transit service	\$2.4 10^{-3} /trip
c'^R_s	Impedance of suburban transit service	\$4.4 10^{-3} /trip

8.2 Numerical results with stationary travel conditions

8.2.1 Aggregate effects of regulation

Table 5 reports the outcome in each regulatory regime under stationary travel conditions for Crowded transit and Diverse transit. The efficiency of the unregulated regime is normalized to 0, and the efficiency of the first-best optimum to 1. The efficiency of the toll (t) or quota (q) is measured using the index

$$\omega \equiv \frac{TC^n - TC^i}{TC^n - TC^f},\ i = t, q,$$

[25]Note that the impedance parameters measure the increase in cost due to each additional traveler; hence their small numerical values.

where TC^n denotes total costs with no regulation, TC^f total costs in the first-best optimum, and TC^i total costs in regulatory regime i, $i = t, q$.

Crowded transit: In the unregulated regime, well over half of each group travels by car. Nearly half of suburban travelers take the bypass, and $N_d^A + N_s^t = 4,107$ travelers out of $10,000$ in total take the direct route. Total costs of $\$208,398$ (not shown) are a little over \$20 per traveler. The second-best optimal toll is \$2.85, and the corresponding second-best optimal quota is $3,070$: about 25% less than usage of the direct route without regulation. Total costs are reduced by $\Delta W = \$2,958$, where ΔW denotes the welfare gain. By comparison, the first-best optimum features a larger transit share for both groups and more use of the direct route by suburban travelers. Total costs decline by $\$5,138$. Thus, the toll or quota achieves a relative efficiency of $\omega = \$2,958/\$5,138 \approx 0.576$.

As shown in row "Prices" of Table 5, if the first-best optimum were decentralized by pricing, tolls and transit fares would be quite high, with a range of roughly \$6 to \$11. With inelastic demand, only the relative prices of the options matter. If the downtown transit fare of \$6.06 is normalized to zero (see row "Differential"), all prices drop by \$6.06. The second-best non-discriminatory toll of \$2.85 is intermediate between the first-best differential tolls of \$1.63 for downtown travelers and \$3.54 for suburban travelers.

Diverse transit: By design, the unregulated equilibrium with Diverse transit is the same as with Crowded transit. The optimal toll of \$4.23 is higher than with Crowded transit because there is no crowding externality, and it is socially efficient to induce a larger transit share. Correspondingly, the optimal quota of $2,569$ is smaller than the optimal quota of $3,070$ for Crowded transit. However, second-best regulation has a much lower relative efficiency of $\omega \approx 0.372$. It performs poorly for two reasons. First, it achieves only 30% of the aggregate increase in transit ridership required to reach the first-best optimum. Second, traffic on the bypass is boosted by more than with Crowded transit, whereas a larger decrease is required. With no transit crowding externality and the same driving congestion externality, the transit share of trips is further below optimal with Diverse transit, and controlling traffic on just one of the four road links is not very effective at redressing the imbalance.

As noted, controlling traffic on the cordon link yields a relative efficiency of about 0.576 with Crowded transit and 0.372 with Diverse transit. We are not aware of any comparable measures of relative efficiency for tradable permits. However, a number of studies have compared the performance of second-best tolls relative to first-best pricing. Some, such as Tabuchi (1993) and de Palma et al. (2005), adopt dynamic models and include transit as an alternative to driving. Others, such as May et al. (2004) and Mun et al. (2003, 2005), use static models without transit. Some feature route choice, while others do not. Overall, studies show that the relative efficiency of tolling depends on a number of factors: the size of the cordon area, level of demand (and hence intensity of congestion), the spatial distribution of trip origins and locations, availability of toll-free alternative routes, and potential to vary toll level by location and time of day. The relative efficiency figures range from about 0.2 to over 0.9, bracketing the values obtained here.

8.2.2 Welfare effects on individual travelers

The welfare effects of the toll and TPS are measured by changes in consumer's surplus (i.e., any payments from selling a permit, minus changes in trip cost). The effects are reported in Table 6. Positive values indicate a gain, and negative values a loss. As Table 2 showed, tolling leaves suburban travelers worse off unless the elastic downtown transit condition is satisfied, which it is not with the base-case parameter values. The welfare effects differ for Crowded transit and Diverse transit.

With Crowded transit, travelers are identical within each group. Tolling leaves downtown travelers \$1.02 worse off, and suburban travelers \$0.29 worse off. With the TPS, downtown travelers still incur a small loss

Table 5: Comparison of unregulated, second-best, and first-best regimes with stationary conditions

Regime	N_d^A	N_d^R	N_s^t	N_s^b	N_s^R	Transit share	Toll	Quota	ΔW	ω
					Crowded transit					
Unregulated	2,240	1,760	1,867	2,774	1,360	31.2%		4,107		0
Second best†	1,815	2,185	1,255	3,320	1,425	36.1%	\$2.85	3,070	\$2,958	0.576
First best:										
Flows	1,474	2,526	1,596	2,646	1,758	42.8%		3,070	\$5,138	1
Prices	\$7.69	\$6.06	\$9.60	\$11.09	\$7.74					
Differential	\$1.63	\$0.00	\$3.54	\$5.03	\$1.67					
					Diverse transit					
Unregulated	2,240	1,760	1,867	2,774	1,360	31.2%		4,107		0
Second best†	1,610	2,390	960	3,584	1,456	38.5%	\$4.23	2,569	\$6,006	0.372
First best:										
Flows	880	3,120	1,511	2,073	2,417	55.4%		2,390	\$16,160	1
Prices	\$3.99	\$0.00	\$5.81	\$8.67	\$0.00					
Differential	\$3.99	\$0.00	\$5.81	\$8.67	\$0.00					

Accounting identities: $N_d^A + N_d^R = N_d$; $N_s^t + N_s^b + N_s^R = N_s$.

† Second-best regulation controls usage of the cordon link, either with a toll or a TPS.

of \$0.14, but suburban travelers gain \$0.59. As noted above, with Diverse transit the toll is higher than with Crowded transit, and the quota is smaller. Nevertheless, downtown travelers in aggregate suffer a smaller average welfare loss of \$0.73 with the toll, and they now gain \$0.36 from the TPS. In aggregate, suburban travelers suffer a slightly higher welfare loss with tolling, but gain a bit more with the TPS. As explained above, travelers in both groups in the (*Transit*, *Transit*) segment neither gain nor lose from tolling. This segment gains the most of the three segments from the TPS. Travelers in the (*Drive*, *Drive*) segment are worse off with tolling. Downtown travelers in this segment also lose from the TPS, while suburban travelers in this segment gain. The welfare effects for the (*Drive*, *Transit*) segment are an unweighted average[26] of the effects for the other two segments.

Overall, for both specifications of transit service and all three traveler segments, suburban travelers fare better from regulation than downtown travelers. There are two reasons. First, suburban travelers have two alternatives to the direct route: the bypass route and transit. Downtown travelers only have transit. Second, suburban travelers pay the same toll as downtown travelers, and face the same permit requirement, despite the fact that suburban travelers create higher external congestion costs (and also higher crowding costs with Crowded transit).

8.3 Numerical results with variable travel conditions

We now turn to the setting of primary interest in which travel conditions vary from day to day. Attention is mainly focused on road capacity shocks. In practice, the probability of shocks depends on three types of factors:[27] environmental factors such as frequency of severe storms, structural characteristics related to road design and hardening, and traffic-related and operational aspects including traffic volume, maintenance, snow clearing, and speed of response to incidents.

[26] The average is unweighted because the distribution of transit costs in the population is assumed to be uniform.
[27] See, for instance, Husdal (2005) and Maoh et al. (2012).

Table 6: Changes in consumers' surplus due to toll and TPS with stationary conditions

Crowded transit						
	Downtown travelers			Suburban travelers		
		Toll	TPS		Toll	TPS
		$-\$1.02$	$-\$0.14$		$-\$0.29$	$+\$0.59$
Diverse transit						
	Downtown travelers			Suburban travelers		
Segment	Fraction	Toll	TPS	Fraction	Toll	TPS
(*Transit*, *Transit*)	0.440	$0.00	$+\$1.09$	0.227	$0.00	$+\$1.09$
(*Drive*, *Drive*)	0.402	$-\$1.51$	$-\$0.43$	0.757	$-\$0.42$	$+\$0.66$
(*Drive*, *Transit*)	0.158	$-\$0.76$	$+\$0.33$	0.016	$-\$0.21$	$+\$0.87$
All	1	$-\$0.73$	$+\$0.36$	1	$-\$0.32$	$+\$0.76$

Changes in consumers' surplus are determined by calculating average private cost for each group and segment without regulation, and subtracting the corresponding cost with regulation.

8.3.1 Individual road link capacity shocks

As an initial exploration into the effects of road-capacity shocks, we consider a scenario in which one road link of the network suffers a 25% reduction in capacity (i.e., a one-third increase in impedance) with a probability 0.1. This scenario would transpire approximately if one lane of a 3-lane highway is closed during half the peak period on average one day every second workweek.[28] Days without a shock will be called "Good days" and denoted by subscript G. Days with a shock will be called "Bad days" and denoted by subscript B. The effects of shocks to each link are summarized in Table 7, first for Crowded transit and then for Diverse transit.

Crowded transit: Row 1 of Table 7 repeats results in Table 5 for the toll and quota without shocks. The fixed toll and fixed quota support the same equilibrium, with no loss relative to flexible instruments. Hence, $\tau^* = \tau^o$, $Q^* = Q^o$, and $E\{TC(Q^*)\} = E\{TC(\tau^*)\} = TC^o$. The equilibrium price of a permit on Good days, q_G, is the same as the toll. The cell for q_B is blank since there are no Bad days.

Shocks to the upper link are shown in row 2. With flexible regulation, on Good days the toll is \$2.85 and the upper cordon carries a flow of $6,390$ vehicles. (We discuss the significance of the flows in this column below.) On Bad days, the flexible toll drops slightly to \$2.74, the flexible quota declines from $3,070$ to $2,718$, and total costs increase by $\Delta TC^o = \$11,539$ compared to Good days. With flexible regulation, expected costs are $(0.1)\$11,539 = \$1,154$ higher than if shocks never occurred.

The fixed toll of \$2.84 is intermediate in value between the flexible tolls for Good and Bad days, although much closer to the toll for Good days because they occur 90% of the time. As indicated in the next column "Loss", total expected costs are barely higher (\$0.44) than with a flexible toll. The fixed quota of $3,034$ is intermediate in value between the flexible quotas for Good and Bad days. The equilibrium permit price on Good days of \$2.95 is slightly higher than the price without shocks of \$2.85 because the fixed quota is below the flexible quota of $3,070$ for Good days. On Bad days, the permit price drops to \$1.84 because the fixed quota is above the flexible quota of $2,718$. This illustrates the potential volatility of permit prices, noted also by Shirmohammadi et al. (2013). The expected loss due to inflexibility of the quota is \$32. This is proportionally

[28] In a previous version of the paper we considered a more severe scenario in which road links suffer a 50% reduction in capacity with a probability 0.2. As expected, the welfare losses from instrument inflexibility were higher than those reported here. However, the qualitative results and welfare rankings of the toll and quota are the same. Moreover, as established in Theorems 3 and 4, expected total costs and the expected losses from inflexibiliity of the instruments vary less than proportionally with the probabilities of shocks.

much higher than the loss for the fixed toll, although barely one-percent of the second-best welfare gain from a flexible quota of \$2,926. Thus, inflexibility of the instruments causes very little welfare loss.

Shocks to the cordon link (row 3) create no welfare loss with a fixed toll since the toll is robust to variations in cordon link capacity, as established in Table 1. The fixed quota of 3,026 is again intermediate between the level of 3,070 on Good days and 2,680 on Bad days and creates an expected welfare loss of \$42. By contrast, the fixed quota results in smaller losses for shocks to the lower link (row 4) and bypass (row 5). Thus, in all cases the losses with a fixed toll and quota are very small.

Diverse transit: Results for Diverse transit, shown in rows 6-10 of Table 7, are similar to Crowded transit. The fixed toll again outperforms the fixed quota, but all the losses are small. The toll performs particularly well for shocks to the upper and cordon links because these shocks increase the private costs of using the direct route, and thus discourage travelers from using it even without intervention. Thus, the flexible toll changes very little. By contrast, since both the private and external costs of using the direct route increase, optimal traffic declines, but with a fixed quota usage does not decrease at all.

The situation differs somewhat for a shock to the bypass because it does not increase the private cost of taking the direct route directly. Demand from suburban travelers to use the direct route increases, and so therefore does the optimal toll. The fixed toll does not help to stem the additional demand, whereas the fixed quota prevents excessive congestion on the direct route. Consequently, the fixed quota performs nearly as well as the fixed toll. A shock to the lower link has mixed effects because it deters suburban travelers from taking the direct route, but does not discourage downtown travelers. Hence, the increase in optimal toll and decrease in optimal quota are intermediate between the changes for the other links.

The literature on network reliability, mentioned in the introduction, has highlighted the importance of critical links that are both vulnerable to shocks and cause significant damage when disruptions occur. In the setting here, vulnerability is quantified by the probability of a capacity reduction, which is fixed at 0.1 for all links. Damages are measured by the increase in traveler costs on Bad days, ΔTC^o. Intuitively, one expects the damages caused by a given percentage capacity reduction of a link to increase with the amount of traffic it carries under normal conditions. As shown in Table 7, this is largely borne out in the example. With both Crowded transit and Diverse transit, the increase in costs has the same rank order as traffic flow on Good days except for the cordon link and bypass, where the order is reversed.

In contrast, traffic flows are a weaker predictor of the welfare loss from inflexiblity of regulation. For the fixed toll, the greatest losses (albeit very small) occur with the bypass, although it carries much less traffic than the upper link. For the fixed quota, the losses on the cordon link are approximately equal to or greater than losses on the upper link although the cordon link carries less than half the traffic.

8.3.2 General road link capacity shocks

Simulations in the previous subsection are limited to capacity reductions on one link at a time. In reality, few links are impervious to shocks and more than one link can be affected at a given time. Shocks to multiple links may be statistically independent or correlated. Independence is plausible in the case of vehicle breakdowns, crashes, road debris, and emergency road repairs. Positive correlation is likely for bad weather that affects driving conditions across a region similarly (Nicholson and Du, 1997). Bad weather can reduce effective capacity by limiting speeds and increasing headways, as well as increasing the frequency of collisions (Dalziell and Nicholson, 2001).

To investigate the effects of multiple capacity shocks, two sets of simulations were conducted: one in which link capacities are perfectly correlated, and the other in which they vary independently. In both cases, each link is assumed to experience a shock with probability 0.1. In the case of correlated shocks, there are only two states

Table 7: Tolls, quotas, and welfare losses from inflexible instruments with individual 25% road capacity shocks

Link with shock	Normal flow	Flexible regulation τ^o	Q^o	ΔTC^o	$E\{\Delta W^o\}$	Fixed toll τ^*	Loss†	Fixed quota Q^*	q_G	q_B	Loss†
					Crowded transit						
1. None		\$2.85	3,070		\$2,958	\$2.85		3,070	\$2.85		
2. Upper	6,390	\$2.74	2,718	\$11,539	\$2,926	\$2.84	\$**0**	3,034	\$2.95	\$1.84	\$**32**
3. Cordon	3,070	\$2.85	2,680	\$3,291	\$2,920	\$2.85	\$**0**	3,026	\$2.97	\$1.76	\$**42**
4. Lower	1,255	\$2.93	2,948	\$861	\$2,960	\$2.86	\$**0**	3,057	\$2.89	\$2.61	\$**4**
5. Bypass	3,320	\$3.02	3,302	\$2,986	\$2,979	\$2.87	\$**1**	3,094	\$2.79	\$3.62	\$**14**
					Diverse transit						
6. None		\$4.23	2,569		\$6,006	\$4.23		2,569	\$4.23		
7. Upper	6,153	\$4.22	2,195	\$7,490	\$6,174	\$4.23	\$**0**	2,530	\$4.34	\$3.27	\$**33**
8. Cordon	2,569	\$4.30	2,220	\$2,281	\$5,954	\$4.24	\$**0**	2,529	\$4.34	\$3.33	\$**32**
9. Lower	960	\$4.37	2,448	\$758	\$6,011	\$4.24	\$**1**	2,557	\$4.26	\$4.06	\$**3**
10. Bypass	3,584	\$4.65	2,736	\$1,493	\$6,092	\$4.27	\$**5**	2,586	\$4.18	\$5.09	\$**7**

† Losses are the difference in expected total costs between inflexible instruments and flexible instruments.

since either all links experience a shock, or none does. With independent shocks, there are $2^4 = 16$ possible states. The probability that no link suffers a shock is $(0.9)^4 = 0.6561$, and the probability that all links suffer a shock is $(0.1)^4 = 0.0001$. Other states occur with intermediate probabilities.

Correlated capacity shocks: For the correlated shocks scenario, links are assumed to experience a 20% reduction in capacity (i.e., an increase of 25% in impedance).[29] The effects are shown in Table 8. With Crowded transit, the fixed toll is set to \$2.86and the fixed quota to 3,016. The welfare loss in either state is negligible for the toll. For the quota, the loss is small without a shock, but \$539 with a shock and the permit price of \$1.60 is \$1.40 below the \$3.00 with no shock. Still, the expected loss of \$61 is only about 2% of the \$2,922 welfare gain from second-best tolling. Overall, the fixed toll again easily outperforms the fixed quota. Results for Diverse transit are similar.

Table 8: Tolls, quotas, and welfare losses from inflexible instruments with correlated 20% road capacity shocks

	Flexible regulation τ^o	Q^o	ΔTC^o	$E\{\Delta W^o\}$	Fixed toll τ^*	Loss†	Fixed quota Q^*	q	Loss†
				Crowded transit					
No shock	\$2.85	3,070			\$2.86	\$**0.02**	3,016	\$3.00	\$**8**
With shock	\$2.94	2,613	\$14,027		\$2.86	\$**2.01**	3,016	\$1.60	\$**539**
Mean	\$2.86			\$2,922		\$**0.22**		\$2.86	\$**61**
				Diverse transit					
No shock	\$4.23	2,569			\$4.27	\$**0.52**	2,512	\$4.39	\$**8**
With shock	\$4.70	2,085	\$8,921		\$4.27	\$**51**	2,512	\$3.28	\$**561**
Mean	\$4.28			\$6,025		\$**6**		\$4.28	\$**64**

† Losses are the difference in expected total costs between inflexible instruments and flexible instruments.

[29] This is at the upper end of observed impacts due to heavy snow and other bad weather conditions.

Independent road capacity shocks: For the independent shocks scenario, the reduction in capacity due to a shock was set to 20%, 33 1/3%, or 50%.[30] Table 9 provides an overview of the effects. As shocks increase in magnitude, the quota ceases to bind in some states: it is "slack". For example, with Crowded transit and 33 1/3% shocks, the quota is slack when all links except the bypass suffer a shock since the direct route is then quite costly for both downtown and suburban travelers. In that state, the unregulated equilibrium number of trips is 2,757: below the quota of 2,979. With Crowded transit and 50% shocks, the quota is slack in seven of the 16 states. As de Palma and Lindsey (2020) note, the possibility that a quota is slack gives a quota more flexibility than a toll. A toll discourages driving by the same amount regardless of conditions, whereas a quota does not discourage driving when driving is sufficiently unattractive anyway. However, this potential advantage of a quota does not materialize in the example. For both Crowded and Diverse transit and all three magnitudes of shocks, welfare losses are much larger for the quota than the toll. Still, the quota can outperform the toll in particular states. For example, in the scenario with Diverse transit and 20% shocks, this happens when the lower link and bypass suffer shocks. These are the links for which the quota performs relatively well in Table 7.

Table 9: Tolls, quotas, and welfare losses from inflexible instruments with independent road capacity shocks

Capacity reduction	$E\{\Delta W^0\}$	Fixed toll τ^*	Fixed toll Average loss†	Fixed quota Q^*	Fixed quota Average loss†	Fixed quota States where slack	Fixed quota States where superior
Crowded transit							
20%	\$2,922	\$2.86	\$**1**	3,018	\$**53**	0	0
33 1/3%	\$2,893	\$2.87	\$**3**	2,979	\$**181**	2	0
50%	\$2,849	\$2.88	\$**9**	3,007	\$**459**	7	0
Diverse transit							
20%	\$6,025	\$4.28	\$**3**	2,514	\$**44**	0	2
33 1/3%	\$6,034	\$4.31	\$**12**	2,462	\$**149**	0	2
50%	\$6,037	\$4.36	\$**35**	2,386	\$**436**	2	1

† Losses are the difference in expected total costs between inflexible instruments and flexible instruments.

8.3.3 Other types of shocks

In addition to road capacities, predictable and unpredictable fluctuations can occur in other model parameters. The effects of such fluctuations are briefly described here. The results suggest that the fixed toll and fixed quota are both relatively robust to these fluctuations.

Variable free-flow link costs: Free-flow travel times can fluctuate due to bad weather, detours, and other reasons. For example, Chung et al. (2006) find that rain reduces median freeway speeds between 4.5 and 8.2 percent. Kyte et al. (2001) quantify how poor visibility, strong winds, wet pavement, and snow all reduce traffic speeds. Bad weather can also increase the frequency of crashes that force traffic to slow down (as well as reduce road capacity). Fluctuations in free-flow costs were examined numerically by assuming that the costs of individual links increase by \$2 or \$4. The welfare losses due to inflexibility are quite small, and the welfare ranking of the fixed toll and fixed quota follows a pattern similar to that for shocks to link capacities.

[30] The corresponding increases in impedance are 25%, 50%, and 100%.

Variable transit capacities: As noted in the introduction, transit service capacity can be reduced by strikes, mechanical failures, and other disruptions. Depending on the cause, shocks may afflict transit lines independently or together. To allow for this, two cases were considered that parallel the treatment of road capacity shocks: one in which downtown and suburban transit service capacities are perfectly correlated, and the other in which they vary independently. In both cases, it is assumed that capacity is reduced by one third with a probability of 0.1. For both Crowded and Diverse transit, the welfare losses are negligible for the fixed toll, and small for the fixed quota. The losses are smaller for independent shocks than correlated shocks.

Variable travel demands: Travel demand typically varies systematically by day of week, as well as unpredictably or irregularly due to weather, major sporting events, and other reasons. To examine the relative performance of the toll and quota numerically, two cases were again considered: one in which downtown and suburban travel demands are perfectly correlated, and the other in which they vary independently. In both cases, demand is assumed to increase by 20% with a probability of 0.2 (e.g., one day per workweek). Similar to the case of transit shocks, the losses from inflexibility are small in all cases (and zero for the toll with Crowded transit, as per Table 1). Again, the losses are smaller for independent shocks than correlated shocks.

8.4 Sensitivity analysis

8.4.1 Alternative parameterizations

The results reported so far are derived with the parameter values shown in Table 4. To assess the sensitivity of results, simulations were performed with other values. The relative performance of fixed tolls and fixed quotas was qualitatively unchanged. However, the efficiency of the two instruments relative to the first-best optimum under stationary conditions varies noticeably. In the base case, the relative efficiency is about 0.576 for Crowded transit, and 0.372 for Diverse transit. The modest performance is attributable to two factors. First, transit ridership can be boosted appreciably only by levying a high toll or imposing a stringent quota. Doing so increases traffic and congestion on the bypass. Second, neither the toll nor the quota discriminates between downtown and suburban drivers despite the fact that suburban drivers create higher external costs.

The relative efficiencies of the toll and quota depend on parameter values. In particular, they increase if parameter values are changed in the following three ways. (1) The fixed costs of transit are reduced. This induces more travelers to take transit so that regulation results in less traffic diversion onto the bypass. (2) The free-flow cost of the bypass is raised. More suburban travelers then use the direct route which is controlled directly. (3) The effective size of the cordon is increased so that fewer travelers can avoid the cordon link if they drive. This can be done numerically by reducing the length of the upper link, increasing the lengths of the other three links, and increasing the number of downtown travelers as a fraction of the total population.[31] By changing parameter values in these ways, the relative efficiencies of the instruments can be increased to over 0.9, similar to the figure obtained by Mun et al. (2003).

8.4.2 Independent capacity shocks with differential probabilities

In Section 8.3.2 it is assumed that shocks are equally likely on all links. Yet, crashes, vehicle breakdowns, road debris, and emergency road repairs tend to be more frequent on long links because there is more opportunity for them to occur.[32] To incorporate this idea in a simple way, the probability of a shock on a link was set

[31] In a real city with geographically dispersed origins and destinations, the optimal cordon is intermediate, rather than maximal, in size; see de Palma et al. (2005) and Mun et al. (2005).

[32] In the case of crashes, frequency also depends on traffic volumes and speed. Allowing for this dependence would complicate the analysis significantly since volumes, speeds, and the probabilities of shocks would all be endogenous, and an iterative process would be required to derive equilibria and optimal regulations.

proportional to its length. The constant of proportionality was chosen so that the average probability of a shock on the four links is the same as in the base case.[33] Simulations were repeated for capacity reductions of 20%, 33 1/3%, and 50%. The results are similar to those in Table 9 with equal probabilities, and to save space are not shown.

8.4.3 Nonlinear congestion costs

The model used so far features linear road link and transit cost functions. Empirical evidence supports linearity in the case of transit crowding costs.[34] Linearity is more questionable for roads. In the Vickrey (1969) bottleneck queuing model, the dependence of trip costs on usage depends on the specification of scheduling preferences. With piecewise-linear preferences, equilibrium trip cost is a linear function of the number of travelers, N (Arnott et al., 1993). With so-called "slope preferences", in which utility of time spent at the origin declines linearly with clock time and utility at the destination increases linearly (Vickrey, 1973), equilibrium cost is a quadratic function of N. With flow, rather than queuing, congestion, the relationship between usage and travel delays is more complex. Traffic engineering studies generally find that delay is a convex function of flow, and many studies still use the US Bureau of Public Roads (1964) travel time function $T = T_0(1 + d(N/K)^\epsilon)$, where T is travel time, T_0 is free-flow travel time, K is a measure of road capacity, d is a positive parameter, and $\epsilon = 4$.

In light of these results, we retain linearity of the crowding cost functions but investigate nonlinear link-cost functions of the form

$$c_a\left(v_a\right) = c_{ao} + c_a' v_a^\epsilon,\ a = u, t, l, b.$$

We reran some simulations numerically with quadratic ($\epsilon = 2$) and quartic ($\epsilon = 4$) functions. The c_a' impedance parameters were rescaled so that the unregulated equilibria with no shocks match the equilibria obtained with linear cost functions.

Table 10 compares the results for individual link capacity shocks in Table 7 and their counterparts with $\epsilon = 2$ and $\epsilon = 4$. Several points are worth noting. First, the expected welfare gain from flexible control, $E\{\Delta W^0\}$, still varies little with the link on which shocks occur. Second, the welfare gain increases with the degree of nonlinearity in congestion. With Crowded transit it increases by over a factor of 5 from about \$3,000 with $\epsilon = 1$ to over \$16,000 with $\epsilon = 4$. With Diverse transit the proportional gain is smaller, but still over 3. The increase is due to the greater marginal external cost of road congestion with higher values of ϵ. Third, the welfare loss from fixed instruments generally increases with ϵ.[35] The loss remains fairly small for the fixed toll, but is appreciable for the fixed quota with $\epsilon = 4$ in the case of shocks to the upper or cordon link. Fourth, in most instances the fixed toll continues to perform better than the fixed quota. The main exception occurs with shocks to the bypass for which the losses are sometimes larger for the fixed toll.

Table 11 compares the results for correlated capacity shocks in Table 8 and their counterparts with $\epsilon = 2$ and $\epsilon = 4$. The pattern is similar to that shown in Table 10. The welfare loss from inflexible instruments again increases with ϵ, and the fixed toll continues to perform better than the fixed quota.

Table 12 compares the results for independent capacity shocks in Table 9 and their counterparts with $\epsilon = 2$ and $\epsilon = 4$ for capacity reductions of 20% and 33 1/3%.[36] Again, the pattern is similar to that in Table 10 and Table 11. The welfare loss from inflexible instruments increases with ϵ, and the fixed toll continues to perform considerably better than the fixed quota. The fixed quota causes large losses with 33 1/3% capacity

[33] A linear relationship between the probability and length clearly cannot hold globally since the probability cannot exceed 1.

[34] Several studies find that crowding costs vary approximately linearly with passenger density. See Whelan and Crockett (2009) and Wardman and Whelan (2011) for the UK, and Haywood and Koning (2015) for Paris.

[35] The two exceptions are with Diverse transit, the toll, and shocks to the lower link and bypass, for which the losses are lower with $\epsilon = 4$ than $\epsilon = 2$.

[36] Simulations with 50% capacity reductions were not performed since the welfare losses with a quota are extreme.

Table 10: Effects of nonlinear congestion costs: individual 25% road capacity shocks

Link with shock	linear ($\epsilon = 1$) $E\{\Delta W^o\}$	Toll Loss	Quota Loss	quadratic ($\epsilon = 2$) $E\{\Delta W^o\}$	Toll Loss	Quota Loss	quartic ($\epsilon = 4$) $E\{\Delta W^o\}$	Toll Loss	Quota Loss
Crowded transit									
Upper	$2,926	**$0**	**$32**	$8,223	**$5**	**$123**	$16,376	**$55**	**$405**
Cordon	$2,920	**$0**	**$42**	$8,229	**$1**	**$101**	$16,547	**$18**	$245
Lower	$2,960	**$0**	$4	$8,351	$1	$8	$16,774	$5	**$15**
Bypass	$2,979	$1	**$14**	$8,460	**$16**	**$19**	$17,072	**$80**	**$32**
Diverse transit									
Upper	$6,174	**$0**	**$33**	$12,064	**$18**	**$112**	$20,467	**$74**	**$358**
Cordon	$5,954	**$0**	**$32**	$12,061	$8	**$72**	$19,381	**$49**	**$749**
Lower	$6,011	$1	$3	$12,206	**$21**	$7	$20,921	$9	**$15**
Bypass	$6,092	$5	$7	$12,429	**$284**	$7	$21,358	**$75**	**$15**

† Losses are the difference in expected total costs between inflexible instruments and flexible instruments.

Table 11: Effects of nonlinear congestion costs: correlated 20% road capacity shocks

State	linear ($\epsilon = 1$) $E\{\Delta W^o\}$	Toll Loss	Quota Loss	quadratic ($\epsilon = 2$) $E\{\Delta W^o\}$	Toll Loss	Quota Loss	quartic ($\epsilon = 4$) $E\{\Delta W^o\}$	Toll Loss	Quota Loss
Crowded transit									
No shock		**$0.02**	**$8.06**		**$0.50**	$29		**$1.86**	$90
With shock		**$2.01**	**$539**		**$54**	**$1,741**		**$219**	**$4,443**
Mean	$2,922	**$0.22**	**$61**	$8,277	**$6**	**$200**	$16,635	**$24**	**$525**
Diverse transit									
No shock		**$0.52**	**$8.39**		**$1.69**	**$27**		**$3.72**	**$75**
With shock		**$51**	**$561**		**$184**	**$1,610**		**$452**	**$3,791**
Mean	$6,020	**$6**	**$64**	$12,197	**$20**	**$185**	$20,854	**$49**	**$447**

† Losses are the difference in expected total costs between inflexible instruments and flexible instruments.

reductions and highly nonlinear ($\epsilon = 4$) congestion costs. In this case, the marginal external cost of congestion on a disrupted link can be very high.

Finally, Table 13 displays detailed results for an intermediate case of Crowded transit with 33 1/3% capacity reductions and $\epsilon = 2$. The fixed toll is set to \$5.40 and the fixed quota to 2,813. The flexible toll ranges from \$4.76 to \$6.72. The fixed toll of \$5.40 is excessive in the five states marked by underlining where it exceeds the flexible toll. The flexible quota ranges from 1,757 to 3,196. The fixed quota of 2,813 is too restrictive in three states, marked by underlining, where it is lower than the flexible quota. Both instruments are too restrictive in the state with no shocks.

Welfare losses from the fixed toll are quite small in all states, with a probability-weighted average of only \$35. Losses are higher with the fixed quota in all states except, marginally, when the lower link and bypass are disrupted simultaneously. The expected loss from the fixed toll is \$521 or 6.3% percent of the welfare gain of \$8,213 from flexible regulation. The quota is slack in two states (i.e., $q = 0$). In these two states and three other states, the efficiency loss exceeds \$5,000. However, all these states occur infrequently.

Table 12: Effects of nonlinear congestion costs: independent road capacity shocks

	linear ($\epsilon = 1$)			quadratic ($\epsilon = 2$)			quartic ($\epsilon = 4$)		
Capacity reduction	$E\{\Delta W^o\}$	Toll Loss	Quota Loss	$E\{\Delta W^o\}$	Toll Loss	Quota Loss	$E\{\Delta W^o\}$	Toll Loss	Quota Loss
Crowded transit									
20%	$2,922	**$1**	**$53**	$8,277	**$11**	**$142**	$16,612	**$57**	**$384**
33 1/3%	$2,893	**$3**	**$181**	$8,213	**$35**	**$521**	$16,422	**$168**	**$3,374**
Diverse transit									
20%	$6,025	**$3**	**$44**	$12,204	**$49**	**$112**	$20,828	**$175**	**$304**
33 1/3%	$6,034	**$12**	**$149**	$12,177	**$69**	**$779**	$20,631	**$207**	**$6,334**

† Losses are the difference in expected total costs between inflexible instruments and flexible instruments.

Table 13: Tolls, quotas, and welfare losses with independent 33 1/3% road capacity reductions (Crowded transit, $\epsilon = 2$)

		Flexible control		Fixed toll	Fixed quota	
Links with shocks	Prob.	τ^o	Q^o	Loss†	q	Loss†
None	0.6561	$\underline{\$5.34}$	$\underline{2,939}$	**$1**	$5.86	**$90**
u	0.0729	$\underline{\$4.76}$	2,245	**$138**	$2.32	**$1,854**
t	0.0729	$5.53	2,371	**$4**	$2.97	**$1,570**
l	0.0729	$5.57	2,768	**$9**	$5.37	**$12**
b	0.0729	$6.29	$\underline{3,196}$	**$223**	$8.04	**$907**
u, t	0.0081	$\underline{\$4.90}$	1,885	**$69**	$0	**$5,547**
u, l	0.0081	$\underline{\$5.02}$	2,047	**$46**	$1.40	**$6,860**
u, b	0.0081	$5.79	2,506	**$39**	$4.24	**$622**
t, l	0.0081	$5.74	2,274	**$29**	$2.48	**$2,428**
t, b	0.0081	$6.59	2,585	**$310**	$5.15	**$450**
l, b	0.0081	$6.43	$\underline{3,018}$	**$284**	$7.41	**$269**
u, t, l	0.0009	$\underline{\$5.13}$	1,757	**$19**	$0	**$5,800**
u, t, b	0.0009	$6.01	2,117	**$81**	$1.35	**$7,011**
u, l, b	0.0009	$5.97	2,316	**$81**	$3.29	**$1,727**
t, l, b	0.0009	$6.72	2,475	**$371**	$4.52	**$1,009**
All	0.0001	$6.18	1,989	**$125**	$0.40	**$7,125**
Mean		$5.42	2,823	**$35**	$5.38	**$521**

u: upper link, t: cordon link, l: lower link, b: bypass

Regulation is too restrictive in the states marked by underlining.

† Losses are the difference in expected total costs between inflexible instruments and flexible instruments.

Figure 3 displays all 16 states in a diagram with traffic volume on the cordon link plotted on the horizontal axis, and the flexible toll on the vertical axis. The fixed toll of $5.40 is depicted by the horizontal dashed line, and the fixed quota of 2,813 by the vertical dashed line. The second-best (flexible) cordon-link volume and toll in each state are identified by a dot. The blue dot in the lower right-hand quadrant depicts Good days in which no shocks occur. The five red dots identify the states in which the welfare loss from the fixed quota exceeds $5,000. In each case, optimal traffic on the cordon link is far below the quota. The two large, unfilled dots identify states in which the quota is slack. The horizontal dotted lines with arrows pointing to the right identify cordon-link volume in these states. For example, in the state shown near the bottom of the figure with shocks to the upper and cordon links, optimal volume is 1,885, and unregulated volume is 2,724: below the quota of 2,813.

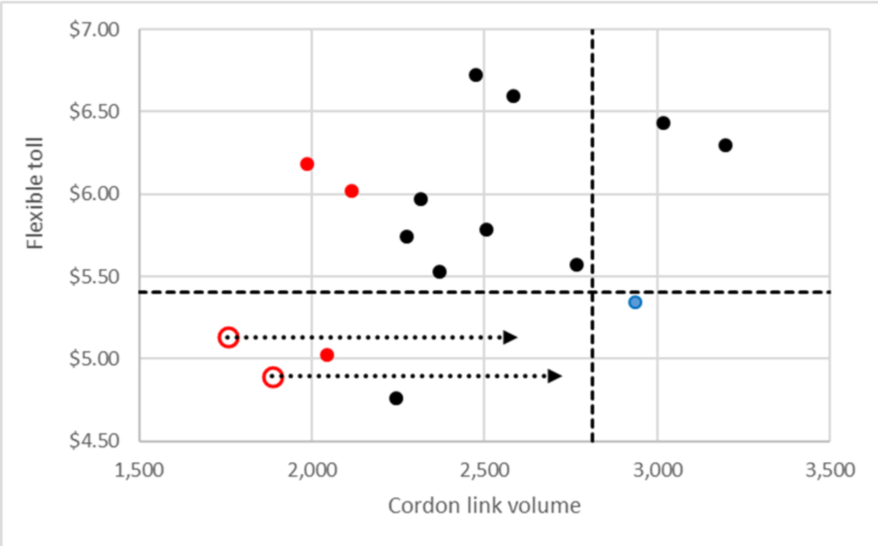

Figure 3: Second-best cordon link volume and flexible toll with independent 33 1/3% capacity reductions (Crowded transit). Each dot depicts one of the 16 states. The blue dot to the lower right identifies Good days. Red dots identify states in which the loss from the fixed quota exceeds $5,000. The two unfilled dots identify states in which the fixed quota is slack.

Results for Diverse transit are broadly similar to Crowded transit, and to save space they are not shown. One difference is that the fixed quota creates a lower welfare loss than the fixed toll in two of the 16 states. Nevertheless, inflexibility of the quota raises the expected costs of shocks 6.4% above the cost with a flexible quota, nearly the same percentage as with Crowded transit.

9 Conclusions

Tolls and tradable permits are alternative tools for tackling road traffic congestion. They are interchangeable if travel conditions are unchanging, but not if conditions vary and the instruments are inflexible. We compare the efficiency of a fixed toll and a fixed permit quota for controlling entry to a downtown area. Travel demand, road conditions, and transit service can all fluctuate. We analyze the model in a series of steps. First, we examine how travelers' mode and route choices vary with each of the model parameters. Second, we assess the robustness of the toll and quota by examining how they vary with parameter values if they could be adjusted freely. With Crowded transit, the flexible toll is invariant to the numbers of travelers in each group and the capacity of the cordon link. Hence, the toll is robust in the sense that inflexibility causes no welfare loss if any or all of these three parameters vary. By contrast, the quota is not robust since the flexible quota varies with all parameter values.

Third, we derive the fixed toll and fixed quota that minimize expected total travel costs. We formalize the idea of a loss function that describes the increase in total travel costs due to inflexibility of either instrument. Using this loss function, we then derive a condition under which one instrument outperforms the other. With Crowded transit, the two instruments perform equally well for fluctuations in free-flow costs of the road links. The toll is robust with respect to fluctuations in capacity of the cordon link, while the quota is not robust. With Diverse transit, the toll is also superior in terms of lower expected total costs for fluctuations in cordon link capacity, whereas the ranking can go either way for capacities of the other links.

Fourth, we use numerical examples to compare the instruments quantitatively in the face of various types of shocks. In most instances, the fixed toll outperforms the fixed quota. In general, the welfare loss from inflexibility of the instruments is quite small as a fraction of the welfare gains from intervention. However, the fixed quota can result in substantial losses in particular states; especially with nonlinear congestion costs. This suggests that the advantage of tradable permits in terms of public acceptability may be at least partly offset by a disadvantage in terms of efficiency. If it is technologically feasible, in such states it might be worthwhile to adjust the quota even if this is inconvenient for travelers. We also compare the welfare-distributional effects of tolls and permits. We assume that travelers do not benefit from the use of toll revenues, and that permits are allocated equally to everyone. We find analytically and numerically that suburban travelers fare better than downtown travelers from both forms of regulation.

Overall, the results show that accounting for network effects is important. The relative efficiency of tolls and permits can depend on which links are most susceptible to shocks. Accounting for network effects is also crucial if travel demand management is imposed on only certain links, as is the case in the model here as well as the cordon-based congestion pricing schemes in operation. Drivers who can avoid traffic control by choosing alternative routes (suburban drivers, here) can cause congestion to be displaced, rather than suppressed. They can also interact with other drivers (downtown drivers, here) with unexpected effects.

The model and the analysis could be extended in various ways. Total travel demand could be made price-sensitive so that tolls and permits have a role to play in controlling numbers of trips as well as mode and route choices. Tolls and permit constraints could be imposed on more of the road network in order to exercise better control of travelers' decisions. Doing so could alter the relative efficiency of the two instruments in the face of shocks. More complex networks and trip-timing decisions could be modeled using a simulator, as done in de Palma et al. (2005) to study cordon and area-based tolls and Rezaeinia (2021) in a preliminary effort to compare tolls and permits with link capacity shocks. Environmental costs of driving could be added. Transit permits could be introduced to internalize crowding externalities and (indirectly) control traffic congestion as well. Finally, health hazards from using transit could be introduced, as in Geng et al. (2021). Health hazards differ from transit crowding in that they are borne partly by travelers, and partly by family members, coworkers, and other people.

Acknowledgments

For helpful comments and suggestions we are grateful to Paolo Delle Site, Patrick Stokkink, and other seminar participants at the CY Transport and Urban Seminar. Thanks are also due to Erik Verhoef and other session participants at the 2021 conference of the International Transportation Economics Association. Any errors are ours. Financial support from the Social Sciences and Humanities Research Council of Canada (Grant 435-2019-0525) is gratefully acknowledged. This study was also supported by the Affinité ANR Research project. Neither institution was involved in any aspect of preparing the paper.

Conflict of Interest

Declarations of interest: none.

References

Agarwal, A. and B. Kickhöfer (2018). The correlation of externalities in marginal cost pricing: Lessons learned from a real-world case study. Transportation 45, 849–873, doi:10.1007/s11116-016-9753-z.

Akamatsu, T. and K. Wada (2017). Tradable network permits: A new scheme for the most efficient use of network capacity. Transportation Research Part C: Emerging Technologies 79.178–195.

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

Belloc, I., J.I. Giménez-Nadal and J. A. Molina Chueca (2022). Weather conditions and daily commuting. IZA Discussion Paper No. 15661 (http://dx.doi.org/10.2139/ssrn.4268769)

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: Policy and Practice 41(7), 672–683.

Brilon, W., J. Geistefeldt and H. Zurlinden (2008). Implementing the concept of reliability for highway capacity analysis, Transportation Research Record 2027, 1-8.

Chen, A., H. Yang, H.K. Lo and W.H. Tang (2002). Capacity reliability of a road network: an assessment methodology and numerical results, Transportation Research B: Methodological 36, 225-252.

Chen, A., C. Yang, S. Kongsomsaksakul and M. Lee (2007). Network-based accessibility measures for vulnerability analysis of degradable transportation networks, Networks and Spatial Economics 7(3), 241-256.

Chung, E., O. Ohtani, H. Warita, M. Kuwahara and H. Morita (2006). Does weather affect highway capacity. Proceedings of the 5th International Symposium on Highway Capacity and Quality of Service, volume 1, 139–146.

Czerny, A.I. (2008). Managing congested airports under uncertainty. In: Czerny, A.I., Forsyth, P. (Eds.), Airport Slots: International Experiences and Options for Reform, Chapter 7, pp. 113–126.

Czerny, A.I. (2010). Airport congestion management under uncertainty. Transportation Research Part B: Methodological 44, 371–380.

Dalziell, E.P. and A.J. Nicholson (2001). Risk and impact of natural hazards on a road network. ASCE Journal of Transportation Engineering 127(2), 159-166.

De Borger, B. and A. Russo (2018). The political economy of cordon tolls. Journal of Urban Economics 105, May 133–148.

de Palma, A., M. Kilani and R. Lindsey (2005). Congestion pricing on an urban road network: A study using the dynamic traffic simulator METROPOLIS. Transportation Research Part A: Policy and Practice 39(7), 588–611.

de Palma, A. and R. Lindsey (1998). Information, welfare and usage of congestible facilities under different pricing regimes. Canadian Journal of Economics 31(3), 666–692.

de Palma, A. and R. Lindsey (2011). Traffic congestion pricing methodologies and technologies. Transportation Research Part C: Emerging Technologies 19(6), 1377–1399.

de Palma, A. and R. Lindsey (2020). Tradable permit schemes for congestible facilities with variable supply and demand. Economics of Transportation 21 100149

de Palma, A., S. Proost, R. Seshadri and M. Ben-Akiva (2018). Congestion tolling — dollars versus tokens: A comparative analysis. Transportation Research Part B: Methodological 108, 261–280.

de Palma, A., S. Vosough and F. Liao (2022). An overview of effects of COVID-19 on mobility and lifestyle: 18 months since the outbreak. Transportation Research Part A: Policy and Practice 159, 372–397.

Delucchi, M.A. and D.R. McCubbin (2003). The health effects of motor vehicle-related air pollution. In D.A. Hensher and K.J. Button (eds.), Handbook of Transport and the Environment. Handbooks in Transport

Volume 4. Elsevier, pp. 411–427.

Dogterom, N., D. Ettema and M. Dijst (2017). Tradable credits for managing car travel: A review of empirical research and relevant behavioural approaches. Transport Reviews 37(3), 322–343.

Dong, J., S. Erdogan, C.-C. Lu and H. S. Mahmassani (2011). State-dependent pricing for real-time freeway management: Anticipatory versus reactive strategies. Transportation Research Part C: Emerging Technologies 19(4), 644–657.

Emmerink, R.H.M., E.T. Verhoef, P. Nijkamp and P. Rietveld (1998). Information effects in transport with stochastic capacity and uncertainty costs. International Economic Review 39(1), 89–110.

Fageda, X., R. Flores-Fillol and B. Theilen (2020). Price versus quantity measures to deal with pollution and congestion in urban areas: A political economy approach. September 7 (https://ssrn.com/abstract=3688332)

Fan, W. and X. Jiang (2013). Tradable mobility permits in roadway capacity allocation: Review and appraisal. Transport Policy 30, 132–142.

Friedrich, R. and E. Quinet (2011). External costs of transport in Europe. In A. de Palma, R. Lindsey, E. Quinet and R. Vickerman (eds.), Handbook in Transport Economics, Cheltenham, UK and Northampton, Mass, USA: Edward Elgar, 369–395.

Gardner, L.M., S.D. Boyles, H. Bar-Gera and K. Tang (2014). Robust tolling schemes for High-Occupancy Toll facilities under variable demand. Transportation Research Record 2450, 152–162.

Gardner, L.M., S.D. Boyles and S.T. Waller (2011). Quantifying the benefit of responsive pricing and travel information in the stochastic congestion pricing problem. Transportation Research Part A: Policy and Practice 45(3), 202–218.

Geng, K., Z. Liang, E.T. Verhoef and Y. Wang (2021), Managing congestion and infection externalities in a road and rail network: pricing versus permits. Paper presented at the Transport and Space-time Economics (TSTE 2021) Academic Clinic, November 23.

Goddard, H.C. (1997). Using tradeable permits to achieve sustainability in the world's large cities: policy design issues and efficiency conditions for controlling vehicle emissions, congestion and urban decentralization with an application to Mexico City. Environmental and Resource Economics 10(1), 63–99.

Grant-Muller, S. and M. Xu (2014). The role of tradable credit schemes in road traffic congestion management. Transport Reviews 34 (2), 128–149.

Grisolía, J.M., F. López and J. de Dios Ortúzar (2015). Increasing the acceptability of a congestion charging scheme, Transport Policy 39, 37–47.

Haywood, L. and M. Koning (2015). The distribution of crowding costs in public transport: New evidence from Paris. Transportation Research Part A: Policy and Practice 77(0), 182–201.

Hörcher, D. and A. Tirachini (2021). A review of public transport economics. Economics of Transportation 25, 100196.

Husdal, J. (2005). The vulnerability of road networks in a cost-benefit perspective, https://husdal.com/2005/01/13/the-vulnerability-of-road-networks-in-a-cost-benefit-perspective/

Jaensirisak, S., M. Wardman and A. D. May (2005). Explaining variations in public acceptability of road pricing schemes. Journal of Transport Economics and Policy 39(2), 127–153.

Jenelius, E., T. Petersen and L.G. Mattsson (2005). Importance and exposure in road network vulnerability analysis. Transportation Research Part A: Policy and Practice 40(7), 537-560.

Job, N., A. van Exel and Piet Rietveld (2001). Public transport strikes and traveller behavior. Transport Policy 8(4), 237-246 DOI: 10.1016/S0967-070X(01)00022-1

Kaddoura, I., L. Kröger and K. Nagel (2016). User-specific and dynamic internalization of road traffic noise exposures. Networks and Spatial Economics. doi:10.1007/s11067-016-9321-2

Kickhöfer, B. and J. Kern (2015). Pricing local emission exposure of road traffic: An agent-based approach. Transportation Research Part D: Transport and Environment 37, 14–28.

Krabbenborg, L., E. Molin, J. A. Annema and B. van Wee (2020). Public frames in the road pricing debate: A Q-methodology study. Transport Policy 93, 46-53, doi:https://doi.org/10.1016/j.tranpol.2020.04.012.

Kyte, M., Z. Khatib, P. Shannon and F. Kitchener (2001). Effect of weather on free-flow speed. Transportation Research Record 1776(1), 60–68.

Lam, W.H.K., H. Shao and A. Sumalee (2008). Modeling impacts of adverse weather conditions on a road network with uncertainties in demand and supply. Transportation Research Part B: Methodological 42, 890–910.

Larcom, S., F. Rauch and T. Willems (2017). The benefits of forced experimentation: Striking evidence from the London underground network. The Quarterly Journal of Economics 132(4), 2019–2055.

Layard, R. (1977). The distributional effects of congestion taxes. Economica 44, 297–304.

Lehe, L. (2019). Downtown congestion pricing in practice. Transportation Research Part C: Emerging Technologies 100, 200–223.

Li, Z. and D.A. Hensher (2010). Toll roads in Australia: An overview of characteristics and accuracy of demand forecasts. Transport Reviews 30(5), 541–569.

Liao, Z., J. Wang, Y. Li and X. Hu (2022). Managing the morning commute problem with tradable credit schemes under a fully autonomous vehicle environment. Systems 10(6), 200 October. DOI: 10.3390/systems10060200

Lindsey, R. (2011). State-dependent congestion pricing with reference-dependent preferences. Transportation Research Part B: Methodological 45(10), 1501–1526.

Lombardi, Claudio, Luís Picado-Santos and Anuradha M. Annaswamy (2021), Model-based dynamic toll pricing: An overview, Applied Sciences 11(11), 4778 DOI: 10.3390/app11114778

Lord, D., A. Manar and A. Vizioli (2005). Modeling crash-flow-density and crash-flow-V/C ratio relationships for rural and urban freeway segments. Accident Analysis and Prevention 37, 185–199.

Maoh, H., B. Anderson and C. Burke (2012). Modeling the resilience of surface freight transportation systems: An application to Ontario, Canada (https://sciencedocbox.com/113330190-Geography/Modeling-the-resilience-of-surface-freight-transportation-systems-an-application-to-ontario-canada.html)

May, A.D., S. Shepherd and A. Sumalee (2004). Optimal locations and charges for cordon schemes. In Santos, G. (Ed.), Road Pricing: Theory and Evidence, Research in Transportation Economics 9, Elsevier Science, pp. 87–105.

Mobility Pricing Independent Commission (2018). Phase 1 Project Update Full Report, January. https://tenyearvision.translink.ca/-/media/tenyearvision/funding/metro_vancouver_mobility_pricing_study.pdf

Mun, S.I., K. Konishi and K. Yoshikawa (2003). Optimal cordon pricing. Journal of Urban Economics 54, 21–38.

Mun, S.I., K. Konishi and K. Yoshikawa (2005). Optimal cordon pricing in a non-monocentric city. Transportation Research Part A: Policy and Practice 39(7-9), 723–736.

Nicholson, A. and Z. Du (1997). Degradable transportation systems: An integrated equilibrium model. Transportation Research Part B: Methodological 31(3), 209–223.

Noland, R.B. and M.A. Quddus (2005). Congestion and safety: A spatial analysis of London. Transportation Research Part A: Policy and Practice 39A, 737-754.

Rezaeinia, P. (2021). Traffic congestion control: A comparative analysis of tradable permits and road tolls. MSc thesis, Sauder School of Business, University of British Columbia.

Rivers, N., S. Saberian and B. Schaufele (2020). Public transit and air pollution: Evidence from Canadian transit strikes. Canadian Journal of Economics 53(2), 496-525, https://doi.org/10.1111/caje.12435

Russo, A., M.W. Adler, F. Liberini and J. N. van Ommeren (2021). Welfare losses of road congestion: evidence from Rome, Regional Science and Urban Economics 89(1), 103692.

Santos, G. and E.T. Verhoef (2011). Road congestion pricing. In A. de Palma, R. Lindsey, E. Quinet and R. Vickerman (eds.), Handbook in Transport Economics, Cheltenham, UK and Northampton, Mass, USA: Edward Elgar, pp. 561–583.

Seshadri, R., A. de Palma and M. Ben-Akiva (2022). Congestion tolling — dollars versus tokens: Within-day dynamics. Transportation Research Part C: Emerging Technologies 143, 103836.

Shirmohammadi, N., M. Zangui, Y. Yin and Y. Nie (2013). Analysis and design of tradable credit schemes under uncertainty. Transportation Research Record: Journal of the Transportation Research Board 2333(1), 27–36.

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

Tabuchi, T. (1993). Bottleneck congestion and modal split. Journal of Urban Economics 34, 414–431.

Taylor, M. and G. D'Este (2007). Transport network vulnerability: A method for diagnosis of critical locations in transport infrastructure systems. In A.T. Murray and T.H. Grubesic (eds), Critical Infrastructure: Reliability and Vulnerability, Advances in Spatial Science, Springer, pp. 9-30.

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

U.S. Bureau of Public Roads (1964). Traffic Assignment Manual, Washington, D.C.: U.S. Bureau of Public Roads.

US FHWA (2020) Transit and Congestion Pricing
(https://ops.fhwa.dot.gov/publications/fhwahop09015/cp_prim7_02.htm)

van Lint, J.W.C., S.P. Hoogendoorn and H.J. van Zuylen (2000). Robust and adaptive travel time prediction with neural networks. In Proceedings of the 6th annual TRAIL Congress, Scheveningen, 12-12-2000 (pp. 1-22). DUP Satellite.

Verhoef, E.T., R.H.M. Emmerink, P. Nijkamp and P. Rietveld (1996). Information provision, flat- and fine congestion tolling and the efficiency of road usage. Regional Science and Urban Economics 26(5), 505–530.

Verhoef, E.T., P. Nijkamp and P. Rietveld (1997). Tradeable permits: Their potential in the regulation of road transport externalities. Environment and Planning B: Planning and Design 24(4), 527–548.

Vickrey, W.S. (1969). Congestion theory and transport investment. American Economic Review (Papers and Proceedings) 59(2), 251-260.

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

Vosough, S., H. Poorzahedy and R. Lindsey (2020). Predictive cordon pricing to reduce air pollution. Transportation Research Part D: Transport and Environment 88, 102564.

Wardman, M. and G. Whelan (2011). Twenty years of rail crowding valuation studies: Evidence and lessons from British experience. Transport Reviews 31(3), 379–398.

Weitzman, M.L. (1974). Prices vs. quantities. The Review of Economic Studies 41(4), 477-491.

Whelan, G.A. and J. Crockett (2009). An investigation of the willingness to pay to reduce rail overcrowding. In: Proceedings of the 2009 International Choice Modelling Conference.

Zhu, S., H. Masud, C. Xiong, Z. Yang, Y. Pan, and L. Zhang (2017). Travel behavior reactions to transit service disruptions: Study of Metro Safetrack projects in Washington, D.C., Transportation Research Record: Journal of the Transportation Research Board 2649, 79–88. http://dx.doi.org/10.3141/2649-09

A Appendix

A.1 Notational glossary

Latin characters

a : arc or link. $a = u, t, l, b$

A : car (automobile) mode

b : bypass link or route.

$c_a(v_a)$: travel cost on link a

c_{a0} : free-flow cost of link a

c'_a : impedance of link a

C_d^A : cost of driving for downtown travelers

C_s^r : cost of driving to the suburb using route r

C_g^R : private cost of rail trip to destination g

C_{g0}^R : fixed cost of rail trip to destination g

$c_g'^R$: impedance of rail service to destination g

$\tilde{C}_g^R$: social cost of rail trip to destination g

d : downtown

f : parameter identifying type of transit ($f = 1$ for Crowded transit, $f = 1/2$ for Diverse transit)

g : group or destination. $g = d, s$

l : lower link

l_a : length of link a

L_τ : welfare loss from imposing fixed toll vs. flexible toll

L_Q : welfare loss from imposing fixed quota vs. flexible quota

m : mode. $m = A, R$

N_g : number of travelers in group g

N_d^A : number of downtown travelers who drive

N_s^r : number of suburban travelers who take route r

N_g^R : number of travelers in group g choosing rail

o : superscript denoting optimal

p : user fee (either toll or permit price)

q : price of a permit

Q : number or quota of permits issued

r : route for suburban travelers. $r = t, b$

R : transit mode

s : suburban

t : cordon link

u : upper link

v_a : traffic volume on link a

x : state

Greek characters

τ : toll

ω : relative efficiency of regulatory regime

A.2 Comparative statics properties of equilibrium flows

Table A.1: Comparative statics properties of equilibrium flows with no regulation or an exogenous toll

Parameter	N_d^A	N_d^R	N_s^t	N_s^b	N_s^R	$v_t = N_d^A + N_s^t$
τ	$-$	$+$	$-Sgn\left[c_d'^R + c_s'^R\right]$	$+$	$Sgn[c_d'^R c_b' - c_u' c_l']$	$-$
N_d	$Sgn[c_d'^R]$	$+$	$-Sgn[c_d'^R]$	$Sgn[c_d'^R(c_s'^R c_t' - c_u' c_l')]$	$Sgn[c_d'^R]$	$Sgn\left[c_d'^R\right]$
N_s	$-Sgn[c_s'^R]$	$Sgn[c_s'^R]$	$Sgn[c_s'^R]$	$Sgn[c_s'^R]$	$+$	$Sgn\left[c_s'^R(c_d'^R c_b' - c_u' c_l')\right]$
c_{u0}	$-$	$+$	$Sgn[c_s'^R c_t' - c_d'^R c_b']$	$-$	$+$	$-$
c_{t0}	$-$	$+$	$-Sgn\left[c_d'^R + c_s'^R\right]$	$+$	$Sgn[c_d'^R c_b' - c_u' c_l']$	$-$
c_{l0}	$+$	$-$	$-$	$Sgn\left[c_d'^R + c_s'^R\right]$	$+$	$-$
c_{b0}	$-Sgn[c_s'^R c_t' - c_u' c_l']$	$Sgn[c_s'^R c_t' - c_u' c_l']$	$Sgn\left[c_d'^R + c_s'^R\right]$	$-$	$+$	$+$
c_u'	Uniform: $-$	Uniform: $+$	Uniform: 0	Uniform: $-$	Uniform: $+$	Uniform: $-$
c_t'	Uniform: $-$	Uniform: $+$	Uniform: 0	Uniform: $+$	Uniform: $-$	Uniform: $-$
c_l'	Uniform: $+$	Uniform: $-$	Uniform: $-$	Uniform: 0	Uniform: $+$	Uniform: 0
c_b'	Uniform: $+$	Uniform: $-$	Uniform: 0	Uniform: $-$	Uniform: $+$	Uniform: $+$
C_{d0}^R	$+$	$-$	$-$	$Sgn[c_s'^R c_t' - c_u' c_l']$	$+$	$+$
C_{s0}^R	$-$	$+$	$+$	$+$	$-$	$Sgn[c_d'^R c_b' - c_u' c_l']$
$c_d'^R$						
$c_s'^R$						

Accounting identities: $N_d^A + N_d^R = N_d$; $N_s^t + N_s^b + N_s^R = N_s$. Cells with complicated formulas are left blank. "Uniform" refers to Uniform transit.

A.3 Comparative statics properties of the flexible toll

Consider the first-order condition for the toll defined in Eq. (9):

$$D_\tau \equiv \frac{dTC\left(\mathbf{N}\left(\tau, x\right), x\right)}{d\tau} = 0.$$

Since the first-order condition holds before and after a change of state,

$$\frac{d}{dx} D_\tau \left(\mathbf{N}\left(\tau, x\right), x\right) = 0.$$

Using the chain rule, one obtains

$$\left[\left(\frac{\partial D_\tau}{\partial \mathbf{N}}\right)^T \frac{\partial \mathbf{N}}{\partial \tau}\right] \frac{\partial \tau^o}{\partial x} = -\frac{\partial D_\tau}{\partial x} - \left(\frac{\partial D}{\partial \mathbf{N}}\right)^T \frac{\partial \mathbf{N}}{\partial x}, \tag{A.1}$$

where ^{T} denotes the transpose. At the optimum, the expression in square brackets on the LHS must be positive by the second-order condition for cost minimization. Hence,

$$\frac{\partial \tau^o}{\partial x} \stackrel{s}{=} -\frac{\partial D_\tau}{\partial x} - \left(\frac{\partial D}{\partial \mathbf{N}}\right)^T \frac{\partial \mathbf{N}}{\partial x}, \tag{A.2}$$

where $\stackrel{s}{=}$ means has the same sign as. The first term on the RHS of Eq. (A.2) corresponds to the direct effect of a parameter change on the first-order condition. The second term corresponds to changes induced by behavioral adjustments. Each term can be positive, negative, or zero.

A.4 Explanations for comparative statics properties with Uniform transit

Table 1 presents the comparative statics properties of the optimal flexible toll. Some of the signs are intuitive. The toll does not depend on the numbers of travelers because, with Uniform transit, transit can accommodate all additional travelers at a constant cost. Hence, road traffic volumes and congestion are unaffected by the numbers of travelers. The toll decreases with the free-flow travel cost on the cordon link (c_{t0}) since fewer drivers use it, and it becomes less congested. Similarly, the toll decreases with the free-flow cost of the upper link (c_{u0}). It increases with the cost of downtown transit (C_{d0}^R) since more downtown travelers choose to drive. Some of the other signs are less obvious:

(a) The toll does not vary with capacity of the lower link, c_l'

This result follows from three observations. First, with Uniform transit, taking transit creates no externalities. Second, as proved in the next section of the appendix, none of the traffic congestion externalities depends on c_l'. Consequently, no externalities on the network are affected by a change in c_l'. Third, none of the derivatives $d\mathbf{N}/d\tau$ in Eq. (A.1) governing the toll depends on c_l'. Hence, the toll itself does not depend on c_l'.

(b) The toll does not vary with capacity of the cordon link, c_t'

The explanation is similar to that for the lower link. No externalities are affected by a change in c_t'. Unlike with the lower link, the derivatives $d\mathbf{N}/d\tau$ do depend on c_t'. However, they all vary proportionally when c_t' changes. Hence, the toll does not depend on the capacity of the very link on which it is imposed.

(c) The toll decreases with the cost of suburban transit, C_{s0}^R

This result may seem paradoxical since an increase in the cost of transit encourages suburban travelers to drive, which contributes to road congestion. To see why the toll drops, note that the cost of driving the direct route and the cost of taking the bypass to the suburbs must both increase, so that suburban traffic rises on each route. Suburban drivers who take the direct route add to traffic on the upper and cordon links, and those taking the bypass add further traffic on the upper link. With Uniform transit to downtown, the cost of driving downtown cannot change. Downtown travellers therefore decrease driving, and by more than the increase in suburban traffic of the direct route. Total traffic on the direct route falls, and therefore so does the optimal toll.

A.5 Proof that with Uniform transit the flexible toll is independent of c_l'

None of the congestion externalities from driving depends on c_l'. This is proved in three steps:

1. With Uniform transit, the cost of driving downtown, $c_{u0}+c_u'v_u+c_{t0}+c_t'v_t+\tau$, is given. Hence, $c_u'v_u+c_t'v_t$ is constant. With linear cost functions, $c_u'v_u+c_t'v_t$ is the external cost of a trip, so the externality caused by driving downtown is constant.
2. With Uniform transit, the cost of using the lower link is fixed at the difference in costs between suburban and downtown transit. Hence, $c_{l0}+c_l'v_l$ is constant, and so is the externality of using the lower link, $c_l'v_l$. The externality caused by taking the direct route to the suburb, $c_u'v_u+c_t'v_t+c_l'v_l$, is therefore constant.
3. With Uniform transit, the cost of taking the bypass route, $c_{u0}+c_u'v_u+c_{b0}+c_b'v_b$, is constant. Hence, the external cost of taking the bypass, $c_u'v_u+c_b'v_b$, is constant, too.

A.6 Derivation of the flexible toll and flexible quota with Crowded transit

Total travel costs are given in Equation (6) in the text:

$$\begin{aligned} TC &= \textstyle\sum_g C_g^R\left(N_g^R\right)N_g^R+\sum_a c_a\left(v_a\right)v_a \\ &= \left(C_{d0}^R+fC_d'^RN_d^R\right)N_d^R+\left(C_{s0}^R+fC_s'^RN_s^R\right)N_s^R+c_u\left(v_a\right)v_u+c_t\left(v_t\right)v_t+c_l\left(v_l\right)v_l+c_b\left(v_b\right)v_b. \end{aligned}$$

With Crowded transit, $f=1$ and

$$TC = N_d(C_{d0}^R+C_d'^RN_d^R)+N_s(C_{s0}^R+C_s'^RN_s^R)-(N_d^A+N_s^t)p,$$

where $p=\tau$ for the toll and $p=q$ for the quota.

The flexible toll is derived by solving Equation (9) in the text. The first-order condition is

$$\frac{dTC}{d\tau}=N_dC_d'^R\frac{dN_d^R}{d\tau}+N_sC_s'^R\frac{dN_s^R}{d\tau}-(N_d^A+N_s^t)-\tau\left(\frac{dN_d^A}{d\tau}+\frac{dN_s^t}{d\tau}\right)=0. \tag{A.3}$$

Derivatives of the flows are

$$\frac{dN_d^R}{d\tau}=\frac{C_s'^R\left(t_b'+t_l'\right)+t_l'\left(t_u'+t_b'\right)}{\Delta},$$

$$\frac{dN_d^A}{d\tau}=-\frac{dN_d^R}{d\tau},$$

$$\frac{dN_s^R}{d\tau}=\frac{C_d'^Rt_b'-t_l't_u'}{\Delta},$$

$$\frac{dN_s^t}{d\tau}=\frac{-C_d'^R\left(C_s'^R+t_b'+t_u'\right)-C_s'^Rt_u'}{\Delta},$$

where

$$\Delta \equiv C_d'^R C_s'^R \left(c_t' + c_l' + c_b'\right) + C_d'^R \left(\left(c_u' + c_b'\right)\left(c_t' + c_l'\right) + c_u' c_b'\right)$$
$$+ C_s'^R \left(\left(c_b' + c_l'\right)\left(c_u' + c_t'\right) + c_t' c_u'\right) + \left(c_b' c_t' + c_b' c_u' + c_u' c_t'\right) c_l'.$$

Substituting the traffic derivatives into the first-order condition yields

$$\frac{dTC}{d\tau} = \tau^o \frac{C_s'^R \left(c_u' + c_b' + c_l'\right) + C_d'^R \left(C_s'^R + c_b' + c_u'\right) + c_l' \left(c_u' + c_b'\right)}{\Delta} + N_d \frac{C_d'^R C_s'^R \left(c_b' + c_l'\right) + C_d'^R c_l' \left(c_u' + c_b'\right)}{\Delta}$$
$$+ N_s \frac{C_s'^R C_d'^R c_b' - C_s'^R c_l' c_u'}{\Delta} - \left(N_d^A + N_s^t\right) = 0.$$

The solution is

$$\tau^o = \frac{\begin{array}{c}\left(C_d'^R C_s'^R + \left(C_d'^R + C_s'^R\right) c_u'\right)\left(c_{b0} - c_{t0} - c_{l0}\right) + \left(C_s'^R (c_b' + c_l') + c_b' c_l'\right)(C_{d0}^R - c_{u0} - c_{t0}) \\ + C_d'^R c_b' (C_{s0}^R - c_{u0} - c_{t0} - c_{l0}) + c_u' c_l' (C_{d0}^R - C_{s0}^R + c_{b0} - c_{t0})\end{array}}{2(c_u' + c_b' + c_l' + C_d'^R) C_s'^R + 2(c_l' + C_d'^R)(c_u' + c_b')}.$$

τ^o does not depend on N_d, N_s, or c_t'. Hence, the flexible toll is robust to variations in these parameters.

The flexible quota is derived by solving Equation (11) in the text. The first-order condition is

$$\frac{dTC\left(q\left(Q\right)\right)}{dQ} = 0,$$

or

$$\left(N_d C_d'^R \frac{dN_d^R}{d\tau} + N_s C_s'^R \frac{dN_s^R}{d\tau} - Q\right)\left(\frac{d\left(N_d^A + N_s^t\right)}{d\tau}\right)^{-1} - q\left(Q\right) = 0. \tag{A.4}$$

This equation is a counterpart to (A.3). After some algebra, the equilibrium permit price can be expressed in the form $q\left(Q\right) = a - bQ$, where

$$a = \frac{\begin{array}{c}\left(c_b' C_d'^R - c_u' c_l'\right) B_s^R + \left(\left(c_u' + c_b'\right) C_d'^R + \left(C_d'^R + c_u'\right) C_s'^R\right) \hat{B}_s^b + \left(c_l' \left(c_u' + c_b'\right) + C_s'^R \left(c_l' + c_b'\right)\right) \hat{B}_d^R \\ + C_d'^R C_s'^R \left((c_l' + c_b') N_d + c_b' N_s\right) + C_d'^R (c_u' + c_b') c_l' N_d - C_s'^R c_u' c_l' N_d\end{array}}{(c_u' + c_l' + c_b' + C_d'^R) C_s'^R + (c_l' + C_d'^R)(c_u' + c_b')},$$

$$b = \frac{\Delta}{(c_u' + c_l' + c_b' + C_d'^R) C_s'^R + (c_l' + C_d'^R)(c_u' + c_b')}.$$

Composite parameter b is positive, so the equilibrium permit price is a decreasing function of the quota. The solution to (A.4) is

$$Q^o = \frac{\begin{array}{c}\left(c_b' C_d'^R - c_u' c_l'\right) B_s^R + \left(\left(C_d'^R + c_u'\right) C_s'^R + \left(c_b' + c_u'\right) C_d'^R\right) \hat{B}_s^b + \left(\left(c_l' + c_b'\right) C_s'^R + \left(c_u' + c_b'\right) c_l'\right) \hat{B}_d^R \\ + 2 C_d'^R C_s'^R \left((N_d c_l' + (N_d + N_s) c_b'\right) + 2 N_d (c_u' + c_b') c_l' C_d'^R - 2 C_s'^R c_u' c_l' N_s\end{array}}{2\left(C_d'^R C_s'^R \left(c_t' + c_l' + c_b'\right) + C_d'^R \left(\left(c_u' + c_b'\right)\left(c_t' + c_l'\right) + c_u' c_b'\right) + C_s'^R \left(\left(c_b' + c_l'\right)\left(c_u' + c_t'\right) + c_t' c_u'\right) + \left(c_b' c_t' + c_b' c_u' + c_u' c_t'\right) c_l'\right)}.$$

The denominator is an increasing function of c_t' and the numerator does not depend on c_t'. Hence, the flexible quota decreases with impedance of the cordon link. It increases with N_d, and increases with N_s if $c_d'^R c_b' - c_u' c_l' > 0$.

A.7 Proof of Theorem 1

Let x' be a state such that $Q^o\left(x'\right) = Q^*$. (Note that x' need not be in the support of $F(\cdot)$.) Let $E\left\{TC_\tau\left(\tau\right)\right\}$ denote expected total cost given the fixed toll of τ, and $E\left\{TC_Q\left(Q\right)\right\}$ expected total cost given a quota Q.

Since τ^* is optimal given $F(\cdot)$,

$$E\left\{TC_\tau\left(\tau^*\right)\right\} \leq E\left\{TC_\tau\left(\tau^o\left(x'\right)\right)\right\}. \tag{A.5}$$

Furthermore,

$$\begin{aligned} &E\left\{TC_\tau\left(\tau^o\left(x'\right)\right)\right\} - E\left\{TC_Q\left(Q^*\right)\right\} \\ &= E\left\{TC\left(\tau^o\left(x'\right), x\right) - TC\left(q\left(Q^o\left(x'\right), x\right), x\right)\right\} < 0, \end{aligned} \tag{A.6}$$

where the inequality follows by assumption. Together, Eqs. (A.5) and (A.6) yield

$$E\left\{TC_\tau\left(\tau^*\right)\right\} < E\left\{TC_Q\left(Q^*\right)\right\}.$$

A.8 Proof of Theorem 3

Expected costs given CDF F_1 are lower with fixed instrument $y^*(F_1)$ than with $y^*(\hat{F})$:

$$ETC^*(F_1) \leq E_{F_1}\left\{TC\left(x, y^*(\hat{F})\right)\right\}.$$

The same is true given CDF F_2:

$$ETC^*(F_2) \leq E_{F_2}\left\{TC\left(x, y^*(\hat{F})\right)\right\}.$$

Hence

$$\begin{aligned} &\lambda ETC^*(F_1) + (1-\lambda)ETC^*(F_2) \\ &\leq \lambda E_{F_1}\left\{TC\left(x, y^*(\hat{F})\right)\right\} + (1-\lambda)E_{F_2}\left\{TC\left(x, y^*(\hat{F})\right)\right\} = E_{\hat{F}}\left\{TC\left(x, y^*(\hat{F})\right)\right\}. \end{aligned}$$

A.9 Proof of Theorem 4

Let $\overline{y}$ be a fixed instrument and $y(x)$ the optimal choice of y in state x. The loss from using $\overline{y}$ in state x is

$$L(x, \overline{y}) = TC(x, \overline{y}) - TC(x, y(x)).$$

The expected loss from using the fixed instrument that is optimal for CDF F is $E_F\left\{L(x, y^*(F))\right\}$.

Again, let $\lambda \in (0,1)$ and define $\hat{F} \equiv \lambda F_1 + (1-\lambda)F_2$. Now

$$E_{F_1}\left\{L(x, y^*(F_1))\right\} \leq E_{F_1}\left\{L(x, y^*(\hat{F}))\right\}.$$

and

$$E_{F_2}\left\{L(x, y^*(F_2))\right\} \leq E_{F_2}\left\{L(x, y^*(\hat{F}))\right\}.$$

Hence

$$\begin{aligned} &\lambda E_{F_1}\left\{L(x, y^*(F_1))\right\} + (1-\lambda)E_{F_2}\left\{L(x, y^*(\hat{F}))\right\} \\ &\leq \lambda E_{F_1}\left\{L(x, y^*(\hat{F}))\right\} + (1-\lambda)E_{F_2}\left\{L(x, y^*(\hat{F}))\right\} = E_{\hat{F}}\left\{L(x, y^*(\hat{F})\right\}. \end{aligned}$$

A.10 Parameterization of the numerical examples

Free-flow costs of road links: The free-flow cost of traversing link a is computed using the formula

$$c_{a0} = \left(\frac{vot^A}{sp_a} + op \right) leng_a,\ a = u, t, l, b,$$

where vot^A is the value of travel time by auto, sp_a is the speed limit on link a, op is vehicle operating cost per km, and $leng_a$ is the length of link a. Values for these and other parameters are given in Table A.2.

Fixed cost of transit: The fixed cost includes the penalty plus the time cost:

$$C_{g0}^R = 8 + \left(\frac{vot^R}{sp^R} \right) dist_g,$$

where vot^R is the cost of travel time by transit, sp^R is the speed of transit, and $dist_g$ is the distance by transit to destination g.

Impedance of road links: Link impedances were chosen by trial and error to obtain plausible modal splits and travel times for all regulatory regimes and parameter variations.

Impedance of transit service: For Crowded transit, $c_d'^R$ and $c_s'^R$ are assumed to be proportional to travel distance. After experimentation, the constant of proportionality was set to 0.0002. Thus, the cost of a transit trip increases by \$0.0002 for each additional km traveled when one more passenger uses the transit line. The same values were used for Diverse transit to facilitate comparisons with Crowded transit.

Table A.2: Parameterization of the numerical examples

Links	Free-flow speed (sp) [km/h]	Length ($leng$) [km]
Upper	60	6
Cordon	50	6
Lower	50	8
Bypass	50	21
Downtown transit	40	12
Suburban transit	40	20
Traveler costs		
Unit cost of travel time by car (vot^A)		\$12/h
Unit cost of travel time by transit (vot^R)		\$18/h
Vehicle operating cost		\$0.15/km
Fixed penalty of using transit		\$8