Tradable permit schemes for congestible facilities with uncertain supply and demand

André de Palma* and Robin Lindsey†

December 3, 2019

Abstract

It is well known that price and quantity regulation are not equivalent under uncertainty. This asymmetry has been a factor in the debate about whether to use taxes or Tradable Permit Schemes (TPS) for controlling greenhouse gas emissions. We analyze the allocative efficiency of a TPS for a congestible facility such as an airport, a road, a recreational area, or a museum that experiences supply and demand shocks. The number of permits issued cannot depend on the state. We compare the efficiency of a TPS and a congestion fee when the level of the fee is similarly constrained to be the same across states. When demand and cost curves are linear, a fee outperforms a TPS for several combinations of additive and/or multiplicative demand and cost shocks. More generally, the ranking depends on the nature and magnitude of demand and cost shocks, the elasticity of the cost function, and whether or not the permit requirement always binds. A TPS tends to perform well when first-best usage levels are similar across states. Analogously, a fee is relatively efficient if first-best fees are similar across states.

Keywords: congestion fee; market efficiency; second-best; tradable permit schemes; uncertainty

JEL Codes: D62; R41; R48

*ENS Paris-Saclay and CREST, Département d'économie et de gestion, Bât. Laplace, 61 Av. du Président Wilson, 94230 Cachan, France. E-mail: andre.depalma@ens-cachan.fr.

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

1 Introduction

Many types of facilities are prone to congestion including roads, airports, seaports, and recreational areas. The estimated costs of congestion delays to consumers, firms, and the overall economy are large.[1] The standard prescription in economics to internalize congestion externalities is a congestion toll or fee. The idea was first proposed for roads by Pigou (1920), and there is now an extensive literature on road-congestion pricing (see Tsekeris and Voß (2008), de Palma and Lindsey (2011), and Santos and Verhoef (2011) for reviews). Marginal-cost pricing of airport congestion was explored by Carlin and Park (1970), and a large body of work on airport congestion fees has developed (see Gillen et al., 2016). Crowding at recreational areas has also been a recurring concern — both at remote areas where solitude is valued highly (e.g., Cicchetti and Smith, 1973; Smith and Krutilla, 1974) and at popular destinations where visitors get in each other's way (Manning, 1999).[2]

Advances in information technology have reduced the costs of imposing congestion-based fees and informing users about fee schedules. However, congestion fees are unpopular. There has been longstanding public and political opposition to road tolls (Jaensirisak et al., 2005) and airport congestion fees (Levine, 2008).[3] In general, fees are also unpopular at recreational areas (Watson and Herath, 1999; Anderson and Freimund, 2004).

In the case of road pricing it is often argued that toll revenues must be dedicated to local transportation to overcome opposition to tolls (Jaensirisak et al., 2005). Yet, if revenues are dedicated to expanding roads or improving public transportation, years may pass after tolling begins before drivers experience any benefits. This creates a chicken-and-egg problem: investments are needed to make tolling acceptable, but revenues from tolls are needed to fund the investments. Moreover, the public may fear project delays and cost overruns, or even that projects will be cancelled — perhaps due to a change of government.[4]

[1] According to Schrank et al. (2017), in 2014 congestion in major US urban areas imposed on drivers approximately 6.9 billion hours of travel delay and 3.1 billion gallons of extra fuel consumption with an estimated total cost of $160 billion. Inrix (2016) reports that traffic congestion in the UK cost motorists more than £30 billion in 2016. According to Ball et al. (2010), in 2007 the costs of air traffic congestion in the United States exceeded $36 billion.

[2] Traditionally, fees at public recreational facilities have been set at modest levels to achieve partial cost recovery while encouraging usage. Managing demand has not been a priority. Nevertheless, in the US there has been slowly growing interest in using fees to redistribute usage over time and space as well as to generate larger revenues. In 1996, the US government enacted a Recreation Fee Demonstration Program to determine the feasibility of using fees to achieve greater cost recovery for operation and maintenance of recreation areas and sites (Espey, 2006). In 2004, the program was replaced by the Federal Lands Recreation Enhancement Act (American Recreation Coalition, 2004).

[3] Objections to road pricing include paying for something that was previously free, double taxation, inequity, and system complexity. See Ecola and Light (2009) and Noordegraff et al. (2014). Commercial airlines and general aviation have long opposed a shift from airport fees based on aircraft weight to landing and takeoff fees based on congestion. In theory, operators with a small share of flights at an airport would pay higher fees than large operators who have an incentive to internalize the costs of delay they impose on their own flights. Price discrimination of this sort is widely viewed as inequitable.

[4] An alternative is to allocate toll revenues to users directly. Kockelman and Kalmanje

Quantity-based tools are an alternative to fees for controlling congestion, pollution and other externalities, and they are widely used. Road travel is rationed by license plate restrictions (Gu et al., 2017), perimeter control (Menelaou et al., 2017), traffic calming (De Borger and Proost, 2013), and other measures (Victoria Transport Policy Institute, 2014). Slot controls are used at airports (Gillen et al., 2016), and quotas are imposed on access to recreational areas (Scrogin, 2005). Quantity controls are generally less efficient than price instruments because they fail to allocate usage to agents who value it the most. However, efficiency can be improved if agents are granted usage rights and allowed to trade them. Slot trading has occurred at US airports since 1986, and in 2008 the European Commission developed guidelines for airport slot trading.[5]

There is now growing interest in the use of Tradable Permit Schemes (TPS) to control road traffic congestion.[6] With a TPS, motorists must acquire a permit to make a trip, traverse a road link, or enter a restricted area — depending on how the TPS is set up. By limiting the number of permits issued, the amount of travel and the resulting congestion can be controlled. Unlike with tolls, if permits are distributed free, drivers in aggregate do not incur an additional net monetary cost to travel. Thus, permits may be able to avoid two common objections to road pricing: that it constitutes double taxation (i.e., taxpayers pay for road construction, and again to use the roads), and that it entails paying for something that was previously free. TPS are often viewed as more equitable, too. It is often claimed that tolling is vertically inequitable because lower-income individuals are willing to pay less for quicker and more reliable trips. This concern is muted with a TPS if permits are given out without charge. Moreover, lower-income individuals and households tend to travel less by car, and hence can earn income by selling excess permits. TPS have the further advantage that motorists gain immediately, rather than having to wait until any promised road or public transport improvements are completed. We further discuss the potential acceptability and equity merits of TPS in the conclusions.

With a few exceptions noted in Section 2, uncertainty has not been considered in weighing the relative merits of congestion fees and quantity controls for congestion management. However, as Weitzman (1974) and others have shown for activities such as pollution control where consumers do not directly bear the costs of market failure, price and quantity regulation are not equivalent if regulatory instruments cannot be adapted to prevailing demand and cost conditions. This has been a factor in the debate about whether a carbon tax or a cap-and-trade scheme is better for controlling greenhouse gas emissions when

(2005) propose that revenues be redistributed monthly to all licensed drivers within an urban region.

[5] See Commission of the European Communities (2008) and Fukui (2010). Auctions are another quantity-based instrument that harnesses the price mechanism. Slot auctions have been implemented at congested airports in Europe. Ball et al. (2006) survey their advantages and disadvantages.

[6] Verhoef et al. (1997) were the first to propose the use of TPS for roads. An extensive literature has now developed; see Fan and Jiang (2013), Grant-Muller and Xu (2014), and Dogterom et al. (2017) for reviews. Most studies are theoretical, although laboratory and field experiments are beginning to be conducted; see Brands et al. (2019).

the rate of climate change and the costs of adaptation and abatement are uncertain. Uncertainty is also relevant for transportation, recreation, and other congestible facilities where demand fluctuates, and capacity or other cost shocks occur. However, the analytics of congestible facilities differ from other goods and services because congestion provides negative feedback to users that limits the size of the deadweight loss due to overusage.

This paper has two main goals. One is to explore the operation of a TPS for congestible facilities when demand and supply conditions are variable, and the quantity of permits issued must be the same regardless of the state. The second is to compare the allocative efficiency of a TPS and a congestion fee when the level of the fee is similarly constrained to be independent of the state. In our model, agents are risk neutral and learn the state before deciding whether to use a facility. Predictable (e.g., seasonal) and unpredictable fluctuations in demand and supply are thus analytically equivalent, and can be treated within a unified framework. We assume that agents are identical other than for their willingness to pay, and consequently we do not compare the welfare-distributional effects of a TPS and a congestion fee. As noted above, a TPS may have a significant advantage over a fee in terms of equity and public acceptability. By ignoring equity, our comparison is arguably biased against TPS. Nevertheless, we find that a TPS can outperform a fee in efficiency under certain plausible conditions.

Our main results are as follows. Similar to Weitzman (1974), we find that a TPS performs well when first-best usage levels are similar across states. Analogously, a fee does well if first-best fees are similar across states. Yet, when we apply Weitzman's assumptions to our model, the results differ sharply. Weitzman showed that with linear demand and cost curves, and additive shocks, quantity control dominates price control if the marginal social cost curve is steeper than the demand curve, and vice versa. By contrast, we show that a fee outperforms a TPS regardless of the slopes of the curves as long as the permit constraint always binds (i.e., the number of permits issued is less than the unregulated level of usage so that permits trade at a positive price). A fee also outperforms a TPS if, in addition to additive shocks, there are multiplicative demand shocks.[7]

These clear-cut results appear to militate against the use of a TPS as an alternative to a congestion fee. Yet, on further exploration, we identify several plausible circumstances in which a TPS can be superior. First, under Weitzman's (1974) assumptions a TPS can be more efficient if the permit constraint does not bind in every state. This happens if demand is sufficiently low, and/or usage costs sufficiently high, that the unregulated equilibrium level of usage is less than the number of permits allows.[8] Total usage with a TPS can there-

[7] Padmanabhan et al. (2010) study multiplicative demand shocks in a model where a profit-maximizing firm decides whether to set a fixed price or a fixed quantity. Shocks to the demand curve can take two forms: horizontal dilations that correspond to variations in the number of consumers, and vertical dilations that correspond to variations in willingness to pay. They show that the firm prefers to set price if horizontal dilations dominate, and to set quantity if vertical dilations dominate.

[8] An instance of a nonbinding permit constraint occurred during the pilot phase of the

fore depend on the state, and this gives a TPS some flexibility in regulating usage. A fixed fee lacks this malleability because users pay the same amount regardless of demand and supply conditions. Second, we show that a TPS can outperform a fee if congestion grows with usage at an increasing rate. If so, it is important to prevent usage from overloading capacity. A permit constraint does this more reliably than a fee. Third, we extend consideration to externalities other than congestion such as pollution, noise, and infrastructure damage. When these additional externalities are present, regulating the amount of usage becomes all-the-more important, and quantity control using a TPS is again favoured over a fee.

The paper is organized as follows. Section 2 reviews the basic results derived by Weitzman (1974) and others on prices versus quantities under uncertainty when, unlike with congestion, consumers do not directly bear the costs of market failure. Section 2 then summarizes what little has been written about congestible facilities. Section 3 presents the model, and describes how a tradable permit system operates. It explains how a TPS can support the first-best optimum when demand and capacity are stationary, and it establishes the equivalence between a TPS and a congestion fee. Section 4 introduces variability in demand and costs, and derives the second-best fixed fee and fixed permit supply. The case of linear demand and cost functions considered by Weitzman (1974) is taken up in Section 5. Section 6 extends consideration to nonlinear functions. A combined fee and TPS scheme is briefly examined in Section 7. Extensions of the model are investigated in Section 8. Section 9 summarizes the main results, and identifies avenues for future research.

2 Prices vs. quantities under uncertainty

In this section we review some established results on prices vs. quantities under uncertainty that will be useful for interpreting and positioning our results. We begin with Weitzman's (1974) classical study.

Weitzman (1974) showed that price and quantity regulation are not equivalent under uncertainty when policy instruments are rigid.[9] He considered a good produced by price-taking firms in a planned economy. Using the notation of our model, total output of the good is N, the marginal benefit (i.e., inverse demand) curve is $p(N)$, and the marginal social cost curve is $MSC(N)$. We use MSC to denote this curve to distinguish it from the private cost curve, $C(N)$, used in our model. MSC includes the full marginal social cost of production which, in Weitzman's model, is borne by firms. In much of the environmental literature, the good is assumed to be pollution abatement and the price of the good corresponds to a tax.[10]

European Union Emissions Trading System when the price of allowances dropped to zero (Merrill Brown et al., 2012).

[9] Rose-Ackerman (1973), Fishelson and Flatters (1975), Fishelson (1976), Adar and Griffin (1976), and Roberts and Spence (1976) independently developed some of Weitzman's ideas.

[10] In this literature it is common to write benefits and costs as functions of the level of

The planner knows only the probability distribution of $p(N)$ and $MSC(N)$ when it makes its decisions, but firms learn the state before making their decisions. Under quantity control, the planner picks a value of N, $\bar{N}$, and firms are obliged to produce $\bar{N}$ regardless of the cost. Under price control, the planner chooses a price, $\bar{p}$. Profit-maximizing price-taking firms produce output N at which $\bar{p} = MSC(N)$. Firms adjust their outputs in response to fluctuations in MSC so that, unlike with quantity control, total output does depend on the state. However, fluctuations in demand alone do not change output because they do not affect either $\bar{p}$ or MSC. Price control and quantity control then support the same (second-best) output and price in all states. If demand variations are uncorrelated with cost variations, only cost uncertainty is relevant to the choice between price and quantity control.

Weitzman takes linear approximations to the demand and MSC curves. He assumes that the slopes of the curves are constant and known, but the planner knows the intercept of the MSC curve only up to a probability distribution. Given additive uncertainty of this sort, Weitzman shows[11] that the difference in expected social surplus between price control and quantity control is equal to

$$\Delta^W = \frac{\sigma_c^2 p'}{2(MSC')^2} + \frac{\sigma_c^2}{2MSC'} = \frac{\sigma_c^2}{2MSC'}\left(1 + \frac{p'}{MSC'}\right), \tag{1}$$

where σ_c^2 is the variance of the intercept of the MSC curve and $'$ denotes a derivative. Price control is superior to quantity control if $\|p'\| < MSC'$ since the expression in brackets is then positive. Quantity control is superior to price control if $\|p'\| > MSC'$.[12] Price control is therefore superior if the demand curve is flatter than the MSC curve, and quantity control is superior if the demand curve is steeper.

Figure 1 depicts an example with two states in which the demand curve is fixed, while the MSC curve is MSC_G in state G (good) and MSC_B in state B (bad). In state G, the optimum is at point g. It can be realized either by setting a quota of N_G^o or fixing the price at p_G^o. In state B, the optimum is at point c. It can be realized either with a quota of N_B^o or a price set at p_B^o. If the quota cannot be differentiated between states, the optimal level, $\bar{N}$, is between N_B^o and N_G^o. Too little output is produced in state G with a deadweight loss equal to area egf. Too much output is produced in state B with a deadweight loss of area cde. If the price cannot be differentiated between states, the optimal fixed price, $\bar{p}$, is between p_G^o and p_B^o. Too much output is produced in state G with a deadweight loss equal to area ghj. Too little output is produced in state B with a deadweight loss equal to area acb.

In the example, output varies too little (i.e., not at all) with quantity control and too much with price control. In both states, the deadweight loss is greater

abatement rather than output. To facilitate later comparison with our model, we instead write them as functions of N.

[11] Weitzman (1974, eqn. (20)).

[12] Adar and Griffin (1976) derive a condition equivalent to (1) using elasticities of the demand and cost curves.

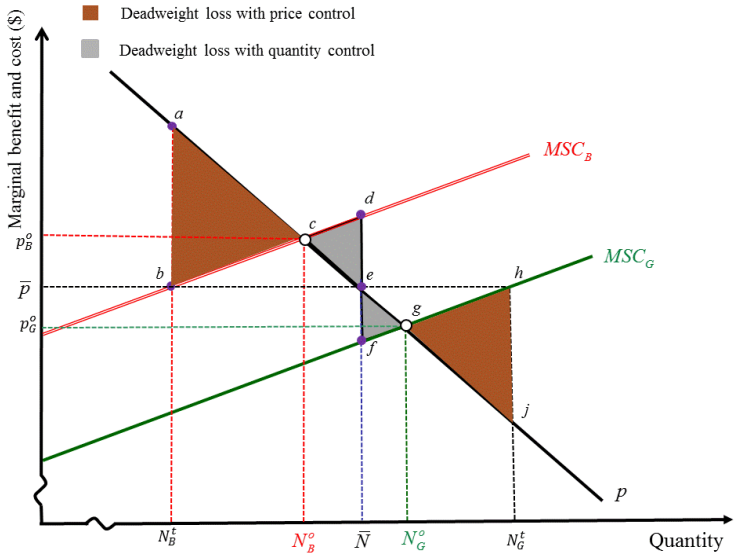

Figure 1: Price and quantity controls in Weitzman's model with variable costs

with price control. Consistent with eqn. (1), quantity control is superior because the demand curve is steeper than the MSC curve. The intuition is straightforward. When the MSC curve is relatively flat, under price control small shifts in the cost curve induce large fluctuations in output. With steep demand, the fluctuations lead either to a serious shortage and a major loss of benefits, or a big surplus and a significant waste of resources. Under these conditions, a fixed quota that guarantees a given output is preferable to a fixed price.[13] By contrast, if the demand curve is relatively flat the exact level of output is not crucial to consumers. And if the MSC curve is steep, forcing production of a given amount can lead either to very high marginal social costs that greatly exceed the benefit, or output at a very low marginal social cost that falls far short of its marginal value to consumers. A fixed price is then superior to a fixed quantity.

Although demand variations alone do not affect the choice between price and quantity control, demand variations do matter if they are correlated with variations in costs. Weitzman (1974) shows that with correlation, eqn. (1) is modified to

$$\Delta^C = \Delta^W - \frac{\sigma_{pc}^2}{MSC'}, \tag{2}$$

where σ_{pc}^2 is the covariance between demand and MSC. Positive correlation favours quantity control, and negative correlation favours price control. As

[13] Following Weyl (2012), imagine that the MSC curve shifts up or down along the steep demand curve. The optimal equilibrium price varies widely whereas the optimal quantity does not, so that quantity control is more efficient.

Stavins (1997) explains, with price control firms respond to high marginal costs by reducing output. If demand is positively correlated with costs, consumption tends to be especially valuable when costs are high so that the reduction in output induced by price control is inappropriate. Quantity control is then more likely to be preferred. The opposite is true with negative correlation.

Weitzman (1974) and most later authors assumed that under price control firms choose the output at which $\bar{p} = MSC(N)$. Laffont (1977) briefly considered another possibility in which consumers decide output by adjusting N so that $\bar{p} = p(N)$. If so, demand variations affect output whereas cost variations do not. This is just the opposite of Weitzman's (1974) case. Intriguingly, Laffont (1977) shows that with additive uncertainty about benefits, the welfare ranking of price and quantity control is given by $-\Delta^W$ in eqn. (1). Quantity control is then superior if the demand curve is flatter than the marginal cost curve, and price control is superior if the demand curve is steeper.

The results summarized thus far apply when the demand and MSC curves are linear and shift vertically. Another possibility is that the curves rotate about fixed intercepts on the vertical axis so that uncertainty is multiplicative rather than additive. Ranking price and quantity control is not as simple in this case. As Laffont (1977) shows, if multiplicative uncertainty is great enough quantity control can outperform price control regardless of whether producers or consumers choose output. Ranking price and quantity control is also not as easy if demand and cost curves are nonlinear (Yohe, 1978).

In summary, the literature suggests that in the face of uncertainty either price control or quantity control can be superior. The welfare ranking depends on the relative slopes of the demand and cost curves and on whether output under price control is decided by producers or consumers. Correlation between demand and costs can also affect the ranking.

For at least two reasons Weitzman's (1974) analysis is not directly applicable to congestible facilities and the choice between fees and a TPS. First, in Weitzman's model consumers do not incur the costs of production. By contrast, the costs of congestion are (largely) borne by users rather than the general population. Congestion provides a negative feedback on usage that limits the costs that congestion can impose. Second, a TPS imposes only an upper bound on usage because not all permits that are issued have to be used. Some may go unused if demand is particularly low, or if (private) costs are particularly high. If so, the TPS will have no effect on usage and the price of permits will drop to zero.[14]

As far as congestible facilities, four studies have analyzed price regulation versus quantity regulation under uncertainty: two on airportion congestion and two on road congestion. Czerny (2008) studies airport congestion using Weitz-

[14] A third distinction between congestible facilities and Weitzman's setting is that most travel — as well as many recreational activities, museum visits, and so on — takes place on networks of multiple congestible links or nodes. To support an efficient distribution of usage over a network using a TPS, upper bounds must be imposed on the flows on each link. This multi-dimensional complication does not arise for a global externality such as greenhouse gas emissions.

man's linear framework and considers separately cases with uncertain demand, uncertain costs, and uncertainy in both demand and costs. His analysis is diagrammatic, and he does not take into account the relationship between private congestion costs and external congestion costs. Czerny (2010) formalizes the analysis in Czerny (2008) algebraically and shows, as we do in Section 5, that with linear functions and additive shocks fees outperform a TPS. Czerny does not consider the possibility that the permit constraint does not bind.

Many studies have examined various aspects of TPS for the use of roads, but only Shirmohammadi et al. (2013) and de Palma et al. (2018) have considered demand and capacity uncertainty. Both papers consider small road networks and use numerical methods to derive solutions. Shirmohammadi et al. (2013) show that the equivalence between TPS and congestion tolls breaks down with uncertainty. They consider the degree of volatility in permit prices, but do not undertake a welfare comparison of TPS and tolls. de Palma et al. (2018) consider a single origin and destination connected by parallel highway routes and a public transport service. Demand for each travel alternative is determined by a mixed-logit choice model. de Palma et al. (2018) solve equilibrium for a large combination of parameter values. They find that a TPS outperforms tolls in a majority of instances although the average difference is not large. As they acknowledge, the complexity of their model makes it difficult to develop intuitive explanations for the results.

Our paper differs from de Palma et al. (2018) in adopting a simpler model with homogeneous agents and a single congestible facility that could be a road, a Central Business District, an airport, a recreational area, a museum, etc.. Agents decide whether to use the facility conditional on the state and the fee or number of permits that are issued. Both fees and permits are constrained to be the same across states. We build on antecedent studies in allowing that the permit requirement may not always bind, and showing how the relative performance of a fee and TPS depends on the nature and magnitude of demand and cost shocks, and the shapes of the demand and cost functions.

Before presenting the model it is worth commenting on three of the assumptions. One is that prospective users know supply and demand conditions when they decide whether to use a facility. Though stringent, this assumption is becoming more plausible as information and communications technology pervades everyday life. For example, drivers can obtain traffic information from traffic websites (e.g., waze.com), GPS devices, connected vehicles, mobile phones, e-mail, social media, and so on. Transit users can get real-time alerts from transit websites as well as mobile apps such as CityMapper, Transit, and Google Maps.

The second assumption is that permit quantities or fees cannot be set according to current information. This might seem inconsistent with the first assumption since facility operators typically have better information than the general public, and may even be a source of public information. The pertinent assumption, however, is that operators cannot use this information to adjust permit quantities or fees on short notice. With the exception of dynamic pricing on some High Occupancy Toll lanes in the US, real-time pricing has not

been used on roads.[15] One reason may be that the infrastructure and operating costs are still too high for it to be cost-effective (Levinson and Odlyzko, 2008). Another is that people prefer simple and predictable pricing schemes (Bonsall et al., 2007). Similar considerations may apply to permits which, so far, have not been implemented for roads, and rarely for other facilities. Transactions costs militate against frequently adjusting the supply of permits, and it seems likely that they would be distributed weekly, monthly, or quarterly rather than daily.

The third assumption is that users are atomistic and disregard any effect they may have on congestion, permit prices, or toll levels. This seems realistic for roads, recreational areas, and museums that numerous people visit per day. In conjunction with the assumption of quasilinear preferences, this allows us to ignore how permits are initially distributed and rule out market power in tradable permits markets.[16]

3 TPS and usage fees: stationary conditions

We begin the analysis by considering a benchmark case in which demand and costs do not vary and are known. Supply conditions are defined by the cost of usage. Consistent with previous studies, we show that a fixed TPS and a fixed usage fee can both support a first-best optimum.

In the model there is a single congestible facility or area, and a continuum of agents who differ only in their willingness to pay to use or access it. The number of agents who decide to use the facility is N.[17] The inverse demand curve is a decreasing and differentiable function, $p(N)$. The cost of usage is $C(N)$. Unless noted otherwise, $C(N)$ is assumed to be strictly increasing and twice continuously differentiable. Thus, the marginal social cost of usage, $MSC(N) = C(N) + C'(N)N$, exceeds $C(N)$. To assure that equilibrium usage is both positive and finite, we adopt:

[15] Elsewhere, toll schedules are adjusted periodically. For example, on SR-91 in Orange County, California, tolls are adjusted every six months to maintain free-flowing conditions on the Express Lanes (https://www.octa.net/91-Express-Lanes/Toll-Policies/, accessed June 1, 2019). In Singapore, toll schedules are adjusted quarterly, and during June and December school holidays, to maintain target speeds on expressways and arterials (https://www.lta.gov.sg/content/ltaweb/en/roads-and-motoring/managing-traffic-and-congestion/electronic-road-pricing-erp.html, accessed June 1, 2019).

[16] Nonatomistic users do exist. Transportation examples include commercial airlines and rail companies, major freight shippers, and even major employers such as government departments. Similarly, large tour companies may generate a substantial fraction of traffic at tourist sites and major recreational areas. He et al. (2013) derive equilibrium conditions for a TPS on a network with nonatomistic users.

[17] A notational glossary is provided in the appendix. We do not model agents' decisions after they have decided to use the facility such as time of use, duration of stay, route, speed of movement, etc.. Yoshimura et al. (2014, 2017) analyze empirically these dimensions of behavior for museum visitors. We also assume that users behave in the same way whether access is unregulated, rationed by price, or rationed by quantity. More et al. (1996) find empirical support for this assumption in the case of campground usage.

Assumption 1 *The demand and cost functions are such that* $p(0) > C(0)$ *and* $p\left(\hat{N}\right) < C\left(\hat{N}\right)$ *for some* $\hat{N} \in (0, \infty)$.

Welfare is measured by total net benefits from usage which equal gross benefits minus total user costs:

$$W(N) = \int_0^N p(n)\, dn - C(N)\, N. \tag{3}$$

3.1 Unregulated equilibrium and first-best optimum

If no fee or TPS is implemented, equilibrium usage, N^n, solves:

$$p(N^n) = C(N^n), \tag{4}$$

where superscript n denotes the no-intervention or unregulated regime. Given Assumption 1, eqn. (4) has a unique and strictly positive solution for N^n.

The first-best optimum (FBO), denoted by superscript o, maximizes net benefits in (3). The first-order condition determining optimal usage, N^o, solves:

$$p(N^o) = C(N^o) + C'(N^o)\, N^o = MSC(N^o). \tag{5}$$

The second-order condition is satisfied if $C''(N^o)\, N^o / C'(N^o) > -2$ (i.e., if the usage cost function is not too concave). Given Assumption 1, it follows that $0 < N^o < N^n < \hat{N}$.

We now consider price and quantity control instruments, beginning with a usage fee since it is more familiar in the literature on congestion management.

3.2 Optimal usage fee

Let N^f denote equilibrium usage when a fee f is levied. The equilibrium condition determining N^f is

$$p\left(N^f\right) = C\left(N^f\right) + f. \tag{6}$$

Comparing eqns. (5) and (6), it is clear that the FBO can be realized by following the standard Pigouvian prescription and charging a fee equal to the marginal external cost (MEC) of usage in the FBO:

$$f^o = C'(N^o)\, N^o = MEC(N^o). \tag{7}$$

The fee in eqn. (7) is unique as long as $C(N)$ is not too concave. If it is highly concave, the inverse demand curve $p(N)$ can intersect the MSC curve more than once. It is then possible (albeit seemingly unlikely) that welfare is maximized at two or more distinct levels of usage supported by different fees. If so, the optimal fee is not unique.

3.3 Optimal tradable permit scheme

A TPS is defined by the total number of permits or credits allocated to potential users, Y, and the number of permits required per use, y. With only one congestible facility, and no distinction between peak and off-peak periods, y can be normalized to 1 so that the maximum permitted usage is Y. On any given day, once permits have been distributed a market opens and agents can buy or sell permits.[18] An equilibrium is assumed to be reached in which the market clears at a price q. Agents are price takers and treat q as given. The full cost of usage is $C(N) + q$. Let N^c denote usage with the TPS. The equilibrium condition for N^c is

$$p(N^c) = C(N^c) + q. \tag{8}$$

If unregulated equilibrium usage exceeds the maximum permitted (i.e., $N^n > Y$), then $q > 0$. If $N^n \leq Y$, then $q = 0$. To support the FBO, Y must be chosen so that $N^c = N^o$. Since $N^o < N^n$, the permit constraint must bind ($N^c = Y$), and permits must trade at a positive price ($q > 0$). Comparing eqns. (8) and (5) it is clear that $q = C'(N^o) N^o$. Equilibrium with the TPS is thus defined by two conditions:

$$Y = N^o, \tag{9}$$

$$q = C'(N^o) N^o. \tag{10}$$

As explained above regarding the fee in eqn. (7), if $C(N)$ is highly concave the optimal level of usage may not be unique. If so, the optimal Y and q will not be unique either.

3.4 Equivalence of a TPS and usage fee

It follows from eqns. (7) and (10) that $q = f^o$: the equilibrium price of a permit matches the Pigouvian fee. It also follows that $qY = f^o N^o$: the total market value of permits matches total usage fee revenues. These results are summarized as:

Proposition 1 *If demand and cost curves are stationary, the FBO can be decentralized using either a fee or a TPS. The equilibrium cost of a permit matches the fee, and the total market value of permits equals total fee revenues.*

Proposition 1 establishes that a TPS and a fee can both support the FBO for a single facility. Yang and Wang (2011) show that the result also holds on a general network with multiple origins, destinations, and links.

[18] Agents will trade unless the number of permits they are initially allocated happens to be commensurate with their individual demands. Individual allocations and willingness to pay can both vary from person to person.

4 TPS and usage fees: Variable conditions

We now turn to the setting of interest in which demand, usage costs, or both, vary over time. Variations are commonplace in the case of road travel. Road capacity is reduced by crashes, slippery conditions, poor visibility, and roadwork. Congestion-free travel time is increased by these factors, as well as by forced traffic diversions to bypass routes that are slower, or less direct, than the preferred route. Costs can vary predictably over time with seasonal variations in fuel prices and vehicle fuel consumption (e.g., higher in very cold weather). Demand fluctuations are also common. Predictable demand fluctuations occur daily, weekly, monthly, and seasonally. Unpredictable fluctuations occur due to inclement weather, special events, transit strikes, and other shocks. Airport congestion is similarly affected by bad weather and demand fluctuations. Demand and capacity variations also occur at outdoor recreational facilities due to weather. Good weather can draw hordes of users. Bad weather can keep people away, while also impairing or preventing usage of such facilities as golf courses, ski slopes, hiking trails, lakeside resorts, beaches and so on.

As in Weitzman's (1974) model, the planner or regulator cannot condition the usage fee or permit allocation on the state. But agents are assumed to learn the state before deciding whether to use the facility. Agents are assumed to be risk neutral so that it does not matter whether variations are predictable or unpredictable. Only the relative frequencies of states matter. Consequently, we will refer to conditions as "variable" rather than "uncertain".

Let Ω denote the set of possible states,[19] and $\omega \in \Omega$ a particular state. Let $p_\omega(\cdot)$ denote the inverse demand function in state ω, $C_\omega(\cdot)$ the cost function in state ω, and N_ω usage in state ω. Finally, let E denote the expectations operator over states. Expected welfare, EW, is given by an extension of (3):

$$EW = E\left\{\int_0^{N_\omega} p_\omega(n)\,dn - C_\omega(N_\omega)\,N_\omega\right\}. \tag{11}$$

In any state ω, unregulated equilibrium usage, N_ω^n, and FBO usage, N_ω^o, can be solved as in Section 3. If the fee and TPS were completely flexible, the FBO could be supported either with a fee $f_\omega^o = C'(N_\omega^o)\,N_\omega^o$ or a permit allocation $Y_\omega = N_\omega^o$. However, by assumption neither the fee nor the permit allocation can be conditioned on the state. Thus, a single value of the fee, f, must be chosen for the pricing instrument. Similarly, a single permit allocation, Y, must be chosen for the quantity instrument. As in Section 3, we first consider the fee and then the TPS.

4.1 Usage control with a fee

With a fee, equilibrium usage in state ω, N_ω^f, is determined by the condition

$$p_\omega\left(N_\omega^f\right) = C_\omega\left(N_\omega^f\right) + f. \tag{12}$$

[19] Set Ω can be continuous or discrete.

The regulator chooses f to maximize expected welfare. Since the term in braces in (11) is continuously differentiable, the derivative and expectations operators can be permuted. The first-order condition for the optimal fee, f^*, is

$$E\left\{\left(p_\omega\left(N_\omega^f\right)-C_\omega\left(N_\omega^f\right)-C'_\omega\left(N_\omega^f\right)N_\omega^f\right)\frac{\partial N_\omega^f}{\partial f}\right\}\Big|_{f^*}=0.$$

Using (12), this condition can be written $E\left\{\left(f^*-C'_\omega\left(N_\omega^f\right)N_\omega^f\right)\frac{\partial N_\omega^f}{\partial f}\right\}\Big|_{f^*}=0$, or

$$\begin{aligned} f^* &= \frac{E\left\{C'_\omega\left(N_\omega^f\right)N_\omega^f\frac{\partial N_\omega^f}{\partial f}\right\}\Big|_{f^*}}{E\left\{\frac{\partial N_\omega^f}{\partial f}\right\}\Big|_{f^*}} \\ &= E\left\{C'_\omega\left(N_\omega^f\right)N_\omega^f\right\}\Big|_{f^*}+\frac{Cov\left(C'_\omega\left(N_\omega^f\right)N_\omega^f,\frac{\partial N_\omega^f}{\partial f}\right)\Big|_{f^*}}{E\left\{\frac{\partial N_\omega^f}{\partial f}\right\}\Big|_{f^*}}, \end{aligned} \tag{13}$$

where Cov denotes covariance. According to the first formula in (13), the optimal fee is a weighted average over states of the marginal external cost (mec) of congestion in each state, $C'_\omega\left(N_\omega^f\right)N_\omega^f$, with weights proportional to the probability of each state and the fee sensitivity of demand in each state, $\partial N_\omega^f/\partial f$.[20] The second formula in (13) shows that the optimal fee exceeds the average mec if the mec is positively correlated with demand sensitivity. Conversely, the fee is less than the mec if the correlation is negative. Unless $N_\omega^f=N_\omega^o$, $C'_\omega\left(N_\omega^f\right)N_\omega^f$ is not equal to the FBO fee in state ω, $C'_\omega\left(N_\omega^o\right)N_\omega^o$. For this reason, we will call $C'_\omega\left(N_\omega^f\right)N_\omega^f$ the *first-best-formula fee*.

Define $\underline{f}^o\equiv Min_\omega f_\omega^o$ as the smallest FBO fee of all states, and $\bar{f}^o\equiv Max_\omega f_\omega^o$ as the largest FBO fee. If $\underline{f}^o<\bar{f}^o$, the first-best fee varies across states. It follows (see the Appendix) that $f^*\in\left(\underline{f}^o,\bar{f}^o\right)$.

Applying the implicit function theorem to (12) one gets $\partial N_\omega^f/\partial f=\left(p'_\omega\left(N_\omega^f\right)-C'_\omega\left(N_\omega^f\right)\right)^{-1}<0$. The weight in state ω is larger the flatter are the usage cost and inverse demand functions because usage is then more sensitive to costs. Substituting this equation into (13), and combining the formula with the bounds on the optimal fee, leads to:

Proposition 2 *Assume demand and cost functions are variable and satisfy Assumption 1. The optimal fixed fee, f^*, solves*

$$f^*=\frac{E\left\{C'_\omega\left(N_\omega^f\right)N_\omega^f\left(C'_\omega\left(N_\omega^f\right)-p'_\omega\left(N_\omega^f\right)\right)^{-1}\right\}\Big|_{f^*}}{E\left\{\left(C'_\omega\left(N_\omega^f\right)-p'_\omega\left(N_\omega^f\right)\right)^{-1}\right\}\Big|_{f^*}}. \tag{14}$$

[20] Equation (13) is analogous to equation (7) in de Palma and Lindsey (1998), who compare state-independent congestion pricing with state-dependent or responsive pricing.

If the FBO fee varies across states, the optimal fixed fee lies strictly within the range of the FBO fee.

As noted in Section 3, if the cost curve is highly concave the optimal fee may not be unique when demand and cost conditions are stationary. The same is clearly true when conditions are variable. Furthermore, with variable conditions the optimal fee can be nonunique even if the cost function is convex. To see this, write the first-order condition for the fee as:

$$E\left\{\frac{f^* - C'_\omega\left(N^f_\omega\right)N^f_\omega}{p'_\omega\left(N^f_\omega\right) - C'_\omega\left(N^f_\omega\right)}\right\}\Bigg|_{f^*} = 0 \tag{15}$$

Suppose eqn. (15) is satisfied with a fee f_1^*. Raising the fee above f_1^* increases the numerator in each state. Since the denominators are all negative, this tends to make the LHS increasingly negative. However, the denominators also change. If the inverse demand curve or cost curve in a state becomes flatter, the denominator shrinks in magnitude and the weight on that state increases. If this trend is concentrated in states where $f^* > C'_\omega\left(N^f_\omega\right)N^f_\omega$, the LHS can change from negative to positive, and another local optimum can be reached. An example is given in the Appendix.[21]

4.2 Usage control with a TPS

Let N^c_ω denote equilibrium usage in state ω with the TPS. As in the stationary setting of Section 3, Y permits are allocated to prospective users. This puts an upper bound of Y on usage in each state:

$$N^c_\omega \leq Y,\ \omega \in \Omega. \tag{16}$$

The TPS operates in the same way as in the stationary setting. In each state, permits are traded freely and the price adjusts so that constraint (16) is satisfied. Let q_ω denote the equilibrium permit price in state ω. The full cost of usage is $C_\omega\left(N^c_\omega\right) + q_\omega$, and the equilibrium condition determining N^c_ω is

$$p_\omega\left(N^c_\omega\right) = C_\omega\left(N^c_\omega\right) + q_\omega. \tag{17}$$

In good states (i.e., in which $p_\omega(\cdot)$ is relatively high, and/or $C_\omega(\cdot)$ is relatively low), the permit constraint (16) binds, $q_\omega > 0$, and $N^c_\omega = Y$. If any states are sufficiently bad, constraint (16) does not bind, $q_\omega = 0$, and $N^c_\omega < Y$.[22] The permit market acts to adjust the monetary price of usage according to demand and cost conditions. This contrasts with the price control scheme in which the monetary price (the fee) is fixed.

[21] The example features linear functions, and is easier to follow after reading Section 5.

[22] If $N^n_\omega = Y$, then $N^c_\omega = Y$ and $q_w = 0$.

Let Ω_C denote the set of states in which the permit constraint binds, and $\Omega_N \equiv \Omega - \Omega_C$ the complementary set (possibly empty) in which the constraint does not bind.[23] It follows that

$$\frac{\partial N_\omega^c}{\partial Y} = \begin{cases} 1 \text{ for } \omega \in \Omega_C \\ 0 \text{ for } \omega \in \Omega_N \end{cases}. \tag{18}$$

The regulator chooses Y to maximize expected welfare given in (11). The first-order condition for the optimal Y, Y^*, is

$$E\left\{ \left(p_\omega \left(N_\omega^c \right) - C_\omega \left(N_\omega^c \right) - C'_\omega \left(N_\omega^c \right) N_\omega^c \right) \frac{\partial N_\omega^c}{\partial Y} \right\} \Bigg|_{Y^*} = 0.$$

Using (17) and (18), this condition reduces to

$$E_{\omega \in \Omega_C} \left\{ q_\omega \right\} |_{Y^*} - E_{\omega \in \Omega_C} \left\{ C'_\omega \left(N_\omega^c \right) N_\omega^c \right\} |_{Y^*} = 0. \tag{19}$$

Similar to the fee, the optimal permit allocation can be bounded above and below. Define $\underline{N}^o \equiv Min_w N_\omega^o$ as the lowest FBO usage level of all states, and $\bar{N}^o \equiv Max_w N_\omega^o$ as the highest level. It follows that $Y^* \in \left[\underline{N}^o, \bar{N}^o \right]$: the optimal permit allocation lies within the range of the FBO usage levels. To see why, note that if $Y < \underline{N}^o$, welfare in every state could be increased by marginally increasing Y. Similarly, if $Y > \bar{N}^o$, welfare in states ω with $N_\omega^n > \bar{N}^o$ could be increased by reducing Y to $\bar{N}^o$ without reducing welfare in other states.

Combining these bounds on the optimal permit allocation with eqn. (19) leads to:

Proposition 3 *Assume demand and cost functions are variable and satisfy Assumption 1. The optimal permit allocation, Y^*, satisfies*

$$E_{\omega \in \Omega_C} \left\{ q_\omega \right\} |_{Y^*} = E_{\omega \in \Omega_C} \left\{ C'_\omega \left(N_\omega^c \right) N_\omega^c \right\} |_{Y^*}. \tag{20}$$

Y^* lies within the range of FBO usage levels of all states.

Unlike Proposition 2 for the fee, usage with the TPS does not necessarily lie *strictly* within the range of first-best usage levels. According to (20), Y^* is chosen so that the expected equilibrium permit price equals the expected marginal external congestion cost. Expectations are only taken over constrained states. Usage conditions in unconstrained states do not affect Y^* because the price of permits is zero in these states and the choice of Y^* does not affect usage. By contrast, the optimal fee (14) does depend on usage conditions in all states. If the permit constraint does not bind in some states, total usage is not completely rigid with a TPS. As will be shown, this flexibility can tip the balance in favour of a TPS.

Two complications arise in applying formula (20). First, the partition of states into Ω_C and Ω_N is not exogenous but depends on Y. Second, if states are

[23] Since function $C(\cdot)$ is strictly increasing, usage is congested in all states including Ω_N.

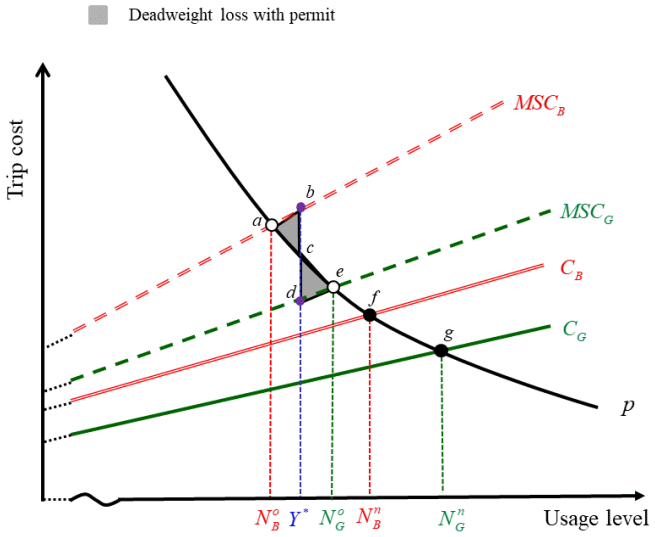

Figure 2: Allocative inefficiency of permits

discrete, the first-order conditions are discontinuous functions of Y, and (20) defines only local optima. Similar to the case with the fee, the optimum Y may not be unique. The two complications are illustrated in Figure 2 using an example with a stationary demand curve and two cost curves corresponding to good (G) and bad (B) usage conditions. The unregulated equilibrium occurs at the point where the demand and cost curves intersect. In state G it is at point g, and in state B at point f. The FBO is found where the demand curve and MSC curve intersect. In state G it is at point e, and in state B at point a.

According to Proposition 3, $Y^* \in [N_B^o, N_G^o]$. Setting $Y = N_G^o$ supports optimal usage in state G. As Figure 2 is drawn, $N_B^o < N_G^o < N_B^n$. Setting $Y = N_G^o$ thus curtails usage in state B too, but not enough to reach N_B^o. Reducing Y slightly below N_G^o reduces overusage in state B further with only a second-order efficiency loss in state G. However, as Y approaches N_B^o the marginal benefit from reducing overusage in state B declines toward zero while the marginal deadweight loss from restricting usage in state G mounts. Hence, given the cost curves in Figure 2, $Y^* \in (N_B^o, N_G^o)$, as shown. Relative to the FBO, the deadweight loss in state G from underusage is measured by the shaded area ced below the inverse demand curve, p, and above the marginal social cost curve, MSC_G. Similarly, the deadweight loss in state B from overusage corresponds to the area abc.

Suppose that the cost curves in the two states differ more sharply than shown in Figure 2 so that $N_B^n < N_G^o$. Reducing Y slightly below N_G^o is now counterproductive because it reduces usage in state G below the FBO level

without affecting usage in state B. Only when Y drops below N_B^n does the permit constraint bind in both states. If the gap between N_B^n and N_G^o is small, and state B is likely enough, restricting usage in state B is optimal, and $Y^* \in (N_B^o, N_G^o)$, again. If the gap is large, it is not worthwhile to set a tight permit constraint that binds in state B, and $Y^* = N_G^o$. If the gap is just the right size, the two options are equally good and the solution is not unique. The optimal permit allocation thus depends, *inter alia*, on the degree of variation in costs and the probabilities of the states.

5 TPS vs. fees: linear functions

We are now ready to tackle the main question addressed in the paper: are prices or quantities more efficient at controlling usage of a congestible facility when conditions vary, and controls cannot depend on the state? To facilitate comparison with the classical results reviewed in Section 2, we assume in this section that the inverse demand and cost curves in each state are linear functions of N:

$$\begin{aligned} p_\omega(N) &= a_\omega - b_\omega N, \\ C_\omega(N) &= c_\omega + d_\omega N, \end{aligned} \tag{21}$$

where parameters a_ω, b_ω, c_ω, and d_ω are all strictly positive. Parameter d_ω governs the rate at which the facility becomes congested, and will be called the *congestion coefficient*. Equilibrium prices, quantities, and welfare depend only on the difference between the intercepts of the inverse demand and cost curves, $A_\omega \equiv a_\omega - c_\omega$. Unless indicated otherwise, we assume $A_\omega > 0$. If state ω is realized, and usage is N, welfare (see eqn. (3)) is

$$W_\omega = A_\omega N - \frac{b_\omega + 2d_\omega}{2} N^2. \tag{22}$$

From eqns. (4) and (5), the unregulated equilibrium and FBO usage levels in state ω are

$$N_\omega^n = \frac{A_\omega}{b_\omega + d_\omega}, \quad N_\omega^o = \frac{A_\omega}{b_\omega + 2d_\omega}. \tag{23}$$

From eqn. (3), the corresponding welfare levels are

$$W_\omega^n = \frac{b_\omega A_\omega^2}{2(b_\omega + d_\omega)^2}, \quad W_\omega^o = \frac{A_\omega^2}{2(b_\omega + 2d_\omega)}. \tag{24}$$

The FBO fee is

$$f_\omega^o = \frac{d_\omega A_\omega}{b_\omega + 2d_\omega}, \tag{25}$$

and the welfare difference between the unregulated equilibrium and the FBO is

$$W_\omega^o - W_\omega^n = \frac{d_\omega^2 A_\omega^2}{2(b_\omega + 2d_\omega)(b_\omega + d_\omega)^2}. \tag{26}$$

The welfare difference is an increasing function of the congestion coefficient d_w. It is a quadratic function of A_ω, and hence of the level of usage in either the unregulated equilibrium or the FBO as per (23). The welfare gain from implementing the FBO fee, the fixed fee, or the TPS all depend on the probability distributions of parameters a_ω, b_ω, c_ω, and d_ω. We consider two cases. The first is Weitzman's case of additive shocks in which the intercept parameters a_ω and c_ω vary, but the slope parameters b_ω and d_ω are constants. In the second case, b_ω and d_ω are variable. We call this case *multiplicative shocks* because the size of the shock is proportional to usage.

5.1 Additive shocks

With additive shocks, a_ω and c_ω are variable, and hence so is A_ω. Define $\bar{A} \equiv E\{A_\omega\}$, and let σ_A^2 denote the variance of A. The optimal fixed fee works out to

$$f^* = \frac{d\bar{A}}{b+2d}, \tag{27}$$

and expected welfare with f^* is

$$EW^{f*} = \frac{\bar{A}^2}{2(b+2d)} + \frac{b}{2(b+d)^2}\sigma_A^2. \tag{28}$$

To analyze the TPS, it is necessary to distinguish between cases in which the permit constraint always binds and cases in which it does not always bind.

5.1.1 Case 1: Permit constraint always binds

Suppose the permit constraint binds in all states so that set Ω_N is empty.[24] Given (17) and (19), the optimal permit allocation is

$$Y^* = \frac{\bar{A}}{b+2d}, \tag{29}$$

and expected welfare is

$$EW^{c*} = \frac{\bar{A}^2}{2(b+2d)}. \tag{30}$$

Given eqns. (28) and (30), the relative advantage of the fee over the TPS is

$$\Delta \equiv EW^{f*} - EW^{c*} = \frac{b}{2(b+d)^2}\sigma_A^2 > 0. \tag{31}$$

Expected welfare with the fee always exceeds expected welfare with the TPS. This result is formalized as:

Theorem 1 *Assume that demand and cost curves are linear, and shocks are additive. If the permit constraint always binds, a fee outperforms a TPS.*

[24] In the Appendix we show that usage with the optimal fee is then always positive so that eqns. (13) and (14) apply.

The ranking in Theorem 1 holds whether or not the demand and cost curves are correlated because expected welfare in eqns. (28) and (30) depends only on the distribution of the difference in their intercepts, $A_w = a_\omega - c_\omega$. Note from eqns. (27) and (29) that $f^* = dY^*$: the optimal fixed fee equals the marginal external cost of congestion evaluated at the optimal permit allocation quantity. In this respect, the two instruments target the same amount of usage although, due to fluctuations in A_w, they support different outcomes.

Theorem 1 differs from the results of Weitzman (1974) and Laffont (1977). Weitzman showed that if firms choose output under price control, price control differs from quantity control only if costs are variable. Price control dominates quantity control if the demand curve is *flatter* than the MSC curve. Laffont showed that if consumers choose output under price control, only demand variations matter. Price control dominates quantity control if the demand curve is *steeper* than the MSC curve. By contrast, Theorem 1 establishes that price control (a fee) dominates quantity control (a TPS) with either cost or demand variability, and regardless of the relative slopes of the demand and cost (or MSC) curves.

Theorem 1 differs from Weitzman (1974) and Laffont (1977) for two reasons. First, users of a congestible facility are both consumers and producers since they incur the costs of usage, not firms. Laffont's (1977) distinction between producers and consumers is thus absent. Second and relatedly, congestion has a negative feedback effect on usage. When costs rise, users curtail their usage somewhat without intervention. Indeed, users in aggregate bear the full costs of congestion, and a market failure exists only because individual users ignore the portion of the costs they generate that are external and imposed on others. The fee is designed to target the externality directly, and if the externality is similar across states a fixed fee can perform nearly as well as the FBO fee. A fixed quota lacks this sensitivity.[25]

Figure 3 illustrates Theorem 1 by presenting an example in which the demand and MSC curves have the same slopes as in Figure 1.[26] In state G, optimal usage is N_G^o and the first-best fee is f_G^o. In state B, the corresponding values are N_B^o and f_B^o. The fee is slightly smaller in state B than state G because usage is lower in state B and (with parallel cost curves) the marginal external cost is lower too. As per eqn. (14), the optimal fixed fee, f^*, is a weighted average of the first-best-formula fees f_G^f and f_B^f. Since $f^* \in (f_B^o, f_G^o)$, the fixed fee supports excessive usage in state G and insufficient usage in state B. The respective deadweight losses in the two states are given by areas hjk and acb. The permit allocation Y^* is set at a level intermediate between N_B^o and N_G^o. The welfare loss is area ehg in state G, and area cde in state B. The qualitative pattern of losses in Figure 3 is the same as in Figure 1, but in both states the losses are much smaller with the fee than the TPS. Thus, in contrast to Figure 1, the fee is superior. Indeed, since the marginal external cost of congestion

[25] As Yohe (1978) notes, quantity constraints also lack this sensitivity in Weitzman's setting.

[26] To make Figure 3 easier to read, curves $C_G(N)$ and $C_B(N)$ are positioned further apart than in Figure 1.

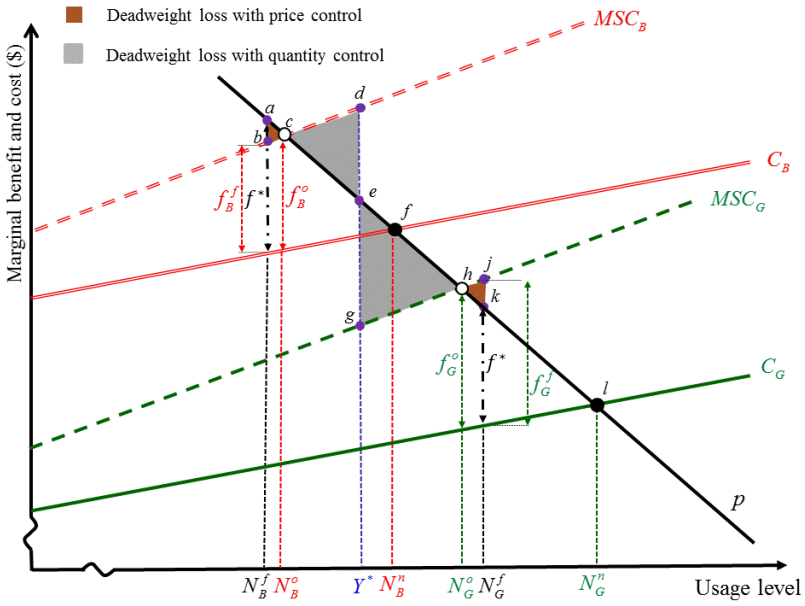

Figure 3: Additive shocks: inverse demand representation

is similar in the two states, so are the Pigouvian fees. The fixed fee is then near-optimal for both states.

Czerny (2010) derives a variant of Theorem 1 in the context of airport congestion for the case where demand is uncertain.[27] He shows how the result can be derived diagrammatically from Weitzman's approach by interpreting the demand curve and cost curve in an appropriate way. He first notes that the net private benefit from usage is measured by the inverse demand minus the user cost: $p(N) - C(N)$. The marginal external cost of usage is the marginal social cost minus user cost: $MEC(N) = MSC(N) - C(N)$. Given (21) and additive shocks, the net benefit and MEC curves are $A_\omega - (b+d)N$ and dN, respectively. The net benefit curve has an absolute slope of $b+d$, and the MEC curve has an absolute slope of d. Since the net benefit curve is steeper, price control dominates quantity control. Theorem 1 generalizes Czerny's result by showing that, if the TPS always binds, price control remains unambiguously superior to quantity control when the intercept parameter c of the cost function is variable, and regardless of how it is correlated with the demand parameter a.

The example in Figure 3 can be depicted using the net benefit and MEC curves as shown in Figure 4. With no regulation, equilibrium usage occurs where the net benefits are zero. Optimal usage is determined where the net benefit curves intersect the MEC curve. The deadweight loss from the fixed fee is small

[27] See his Proposition 1. Czerny does not include an intercept parameter in the cost function so that $c_\omega = 0$ in eqn. (21). Thus, he does not explicitly consider additive cost shocks although, as shown here, his results continue to hold if cost shocks can occur.

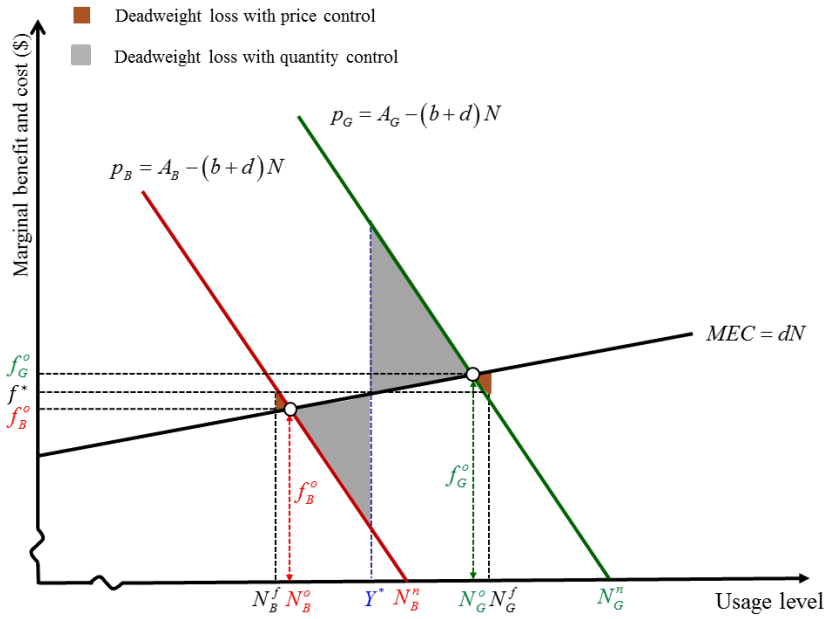

Figure 4: Additive shocks: net demand representation

because the MEC curve does not vary much over the range of optimal usage.

5.1.2 Case 2: Permit constraint does not always bind

If the permit constraint does not bind in some states, set Ω_N is not empty. Given eqns. (22) and (24), expected welfare is

$$EW^c = E_{\omega\in\Omega_C}\{A_\omega\}Y - E\{\Omega_C\}\frac{b+2d}{2}Y^2 + \frac{b}{2(b+d)^2}E_{\omega\in\Omega_N}\{A_\omega^2\}, \quad (32)$$

where $E\{\Omega_C\}$ is the probability that the permit binds. The optimal Y is derived by maximizing eqn. (32). In contrast to the case in which the permit constraint always binds, the TPS can outperform a fee:

Theorem 2 *Assume that demand and cost curves are linear, and shocks are additive. If the permit constraint does not always bind, a TPS can outperform a fee.*

We prove Theorem 2 using an example. Suppose A has a two-point distribution: $A = A_B$ with probability $\wp$, and $A = A_G$ with probability $1-\wp$ where

$A_G > A_B > 0$. The TPS outperforms the fee when[28]

$$\frac{A_B}{A_G} < \frac{\sqrt{1-\wp} - (1-\wp)}{\wp}. \tag{33}$$

Condition (33) is satisfied when the states differ sufficiently. Since the maximum value of the right-hand side is 0.5 (reached in the limit as $\wp \to 0$), the TPS can outperform the fee only when $A_B/A_G < 0.5$. When (33) is satisfied, it is better with the TPS to ignore state B and support the FBO in state G. As explained in discussing Figure 2, the welfare gain from improving usage in state G dominates the loss from doing nothing in state B. The fee is less efficient than the TPS because it does not discriminate between the two states despite their large differences. Condition (33) is notable in that it does not depend on parameters b and d, and therefore does not depend on the relative slopes of the demand and cost curves. This provides another contrast with Weitzman's setting.[29]

5.2 Multiplicative shocks

Suppose now that the intercept parameters a_ω and c_ω are constants, while the slope parameters b_ω and d_ω are variable so that shocks are multiplicative.[30] Multiplicative cost shocks are a natural assumption if capacity is variable because congestion costs are often assumed to depend on the ratio of usage to capacity (and thus be homogeneous of degree zero).[31] Multiplicative demand shocks occur if the number of agents with a given reservation price varies across states by the same proportion at all reservation-price levels.

Applying eqn. (24), expected welfare in the unregulated and FBO regimes with multiplicative shocks is:

$$EW^n = \frac{A^2}{2} E\left\{\frac{b_\omega}{(b_\omega + d_\omega)^2}\right\}, \quad EW^o = \frac{A^2}{2} E\left\{\frac{1}{b_\omega + 2d_\omega}\right\}. \tag{34}$$

[28] See the Appendix for details. In order for usage in state B to be positive with the fee, parameter values must also satisfy $A_B/A_G > \wp d/\left(b + (1+\wp)\, d\right)$. This condition is always satisfied when b is large enough. If this condition does not hold, both the fee and the TPS should be optimized for state G.

[29] Goodkind and Coggins (2015) also study the possibility of corner solutions in Weitzman's model. They consider a polluting industry for which corner solutions arise if there is either no abatement or complete abatement. Interestingly, they find — contrary to the result here — that corner solutions favor price control over quantity control.

[30] Adar and Griffin (1976) consider multiplicative uncertainty in a theoretical pollution-control model. Watson and Ridker (1984) assume multiplicative uncertainty in an empirical study of air and water pollution control, and Hoel and Karp (2001) do likewise in a study of stock pollutants.

[31] If the cost function is linear, it takes the form $C = c + kN/s$, where s is capacity and k is a positive constant. Variations in capacity then translate to variations in the congestion coefficient d in equation (21).

If a fee f is levied, usage is $N_\omega^f = (A-f)/(b_\omega + d_\omega)$, and expected welfare is

$$EW^f = E\left\{\frac{1}{b_\omega + d_\omega} A(A-f) - \frac{b_\omega + 2d_\omega}{2(b_\omega + d_\omega)^2}(A-f)^2\right\}.$$

From the first-order condition $\partial EW^f/\partial f = 0$, the optimal fixed fee is

$$f^* = A\frac{E\left\{\frac{d_w}{(b_\omega+d_\omega)^2}\right\}}{E\left\{\frac{b_\omega+2d_\omega}{(b_\omega+d_\omega)^2}\right\}}.$$

Usage with the optimal fixed fee is strictly positive in all states. Expected welfare is

$$EW^{f*} = \frac{A^2}{2}\frac{\left(E\left\{\frac{1}{b_\omega+d_\omega}\right\}\right)^2}{E\left\{\frac{b_\omega+2d_\omega}{(b_\omega+d_\omega)^2}\right\}}. \tag{35}$$

With multiplicative shocks, demand and cost shocks can no longer be treated jointly using a composite parameter such as A_ω for additive shocks. To see why, note that in Figure 4 a shock to the demand parameter b affects only the net benefit curve whereas a shock to the congestion coefficient d affects both the net benefit curve and the MEC curve.

5.2.1 Multiplicative cost shocks

If only multiplicative cost shocks occur (i.e., only the congestion coefficient d is variable), a fee outperforms a TPS. This result is proved in the Appendix and stated as:

Theorem 3 *Assume that demand and cost curves are linear, and only the congestion coefficient is variable. Then a fee outperforms a TPS.*

Theorem 3 is a counterpart to Theorem 1 which applies for additive demand and cost shocks. Theorem 3 is more limited in that it only applies to cost shocks. However, Theorem 3 is less restrictive in that the permit constraint does not have to bind.

To see why a fee is superior, note that a shock to the congestion coefficient d affects both the net benefit curve and the MEC curve. If d increases, the net benefit curve rotates clockwise downwards while the MEC curve rotates counterclockwise upwards. Both movements reduce optimal usage. The range of optimal usage can vary substantially which militates against a fixed permit quantity. By contrast, the downward shift in the net benefit curve reduces the optimal fee whereas the upward shift in the MEC curve increases it. The net effect on the fee is muted so that a fixed fee performs relatively well. Consistent with eqn. (2) in Weitzman (1974), negative correlation between the demand and cost curves works in favour of price control.

5.2.2 Multiplicative demand shocks

With multiplicative demand shocks it is again possible to derive conditions such that a fee is welfare superior to a TPS. The conditions are more restrictive than for multiplicative cost shocks in that the permit constraint must always bind. However, the conditions are less restrictive in that additive shocks to both demand and costs can also occur. We formalize this result as:

Theorem 4 *Assume that demand and cost curves are linear, the congestion coefficient d is constant, and the permit constraint always binds. Then, with any combination of multiplicative demand shocks and additive demand or cost shocks, a fee outperforms a TPS.*

Theorem 4 is proved in the Appendix. The theorem is significant in two respects. First, since it encompasses not only additive demand and cost shocks but also multiplicative demand shocks it generalizes Theorem 1.[32] The joint probability distribution of parameters a, b, and c is unrestricted so that Theorem 4 covers various types of shocks. In particular, it covers demand shocks that affect the willingness to pay of all users by the same multiplicative factor so that the demand curve rotates about a fixed intercept on the horizontal (quantity) axis.[33] Second, the proof of Theorem 4 entails showing that given *any* permit allocation Y, the fixed fee $\hat{f} = dY$ supports a more efficient usage level than the TPS *in every state*. In particular, the fixed fee $\hat{f} = dY^*$ outperforms the optimal TPS not only in terms of expected welfare, but in every possible state. The fact that a suboptimal fee Pareto dominates the optimal TPS provides a clear sense that the fee is superior.[34]

Theorem 4 does require that the optimal permit constraint always bind. Parameters a, b, and c thus cannot vary too much. If they do fluctuate a lot, so that the constraint does not always bind, a TPS can outperform a fee. This is demonstrated in the Appendix using an example similar to that used to prove Theorem 2.

5.2.3 Multiplicative demand and cost shocks

If multiplicative demand and cost shocks both occur, the TPS and fee cannot be ranked in general. To see this, suppose that A is constant and the permit

[32] It also generalizes Czerny (2010) by extending consideration to both additive cost shocks and multiplicative demand shocks, as well as recognizing that the permit constraint must bind.

[33] The ratio a/b is then the same in all states. This type of shock corresponds to vertical dilations of the demand curve in Padmanabhan et al. (2010). In contrast to Theorem 4, they show that a profit-maximizing firm prefers to set quantity rather than price. Indeed, if production is costless the profit-maximizing quantity depends only on a/b and hence is the same in all states.

[34] Formulas for the optimal f^* and Y^* are given in the Appendix. Unlike with eqns. (27) and (29) for additive shocks, it is not generally true that $f^* = dY^*$ so that the optimal fee does not target the same output as the permit. Also, unlike the fee $\hat{f}$, the optimal fee f^* does not necessarily yield a higher welfare than the TPS in every state.

constraint always binds. The optimal permit allocation is

$$Y^* = \frac{A}{E\{b_\omega + 2d_\omega\}}, \tag{36}$$

and expected welfare is

$$EW^{c*} = \frac{A^2}{2E\{b_\omega + 2d_\omega\}}. \tag{37}$$

Given eqns. (35) and (37),

$$EW^{f*} - EW^{c*} \stackrel{s}{=} \left(E\left\{\frac{1}{b_\omega + d_\omega}\right\}\right)^2 E\{b_\omega + 2d_\omega\} - E\left\{\frac{b_\omega + 2d_\omega}{(b_\omega + d_\omega)^2}\right\}, \tag{38}$$

where $\stackrel{s}{=}$ means has the same sign as. Introducing the composite variables $x_\omega \equiv b_\omega + d_\omega$ and $y_\omega \equiv b_\omega + 2d_\omega$, eqn. (38) can be written

$$EW^{f*} - EW^{c*} \stackrel{s}{=} \left(E\left\{\frac{1}{x_\omega}\right\}\right)^2 E\{y_\omega\} - E\left\{\frac{y_\omega}{x_\omega^2}\right\}. \tag{39}$$

Suppose parameters b and d are perfectly negatively correlated, **and vary** such that variable y_ω is constant. Only variable x then depends on the state, and it follows from Jensen's inequality that (39) is negative. A TPS is then welfare-superior to a fee. If parameters b and d are negatively correlated, demand is high when costs are high, and vice versa.[35] Optimal usage is then insensitive to the state, and a TPS performs well. This is consistent with eqn. (2) and Stavins' (1997) result that positive correlation between demand and costs favours quantity control.

5.3 Additive and multiplicative cost shocks

Theorems 3 and 4 establish that, under relatively general conditions, with either additive or multiplicative cost shocks a fee is welfare-superior to a TPS. This is no longer true if both types of cost shocks occur. Consider the example in Figure 5 featuring two states, G and B. The cost curve in state B has a larger intercept (c) than the cost curve in state G, but a lower slope (d) so that it features a positive additive shock and a negative multiplicative shock.[36] The marginal social cost of usage is lower in state G when usage is light, but higher when usage is heavy. The demand curve crosses the MSC curves where they

[35] Correlation of this sort could be due to weather. For example, in bad weather travelers may prefer to drive rather than walk or take transit, but driving is slowed by poor visibility or slippery conditions. Similarly, after a fresh snowfall skiing conditions are excellent but getting to the ski area can be difficult.

[36] With road transportation this is possible if on bad days drivers are forced off their normal route onto an alternative that is more circuitous, but has a higher capacity. Similarly, in bad weather a narrow, scenic trail may be closed, and hikers redirected to a rougher but wider alternative path.

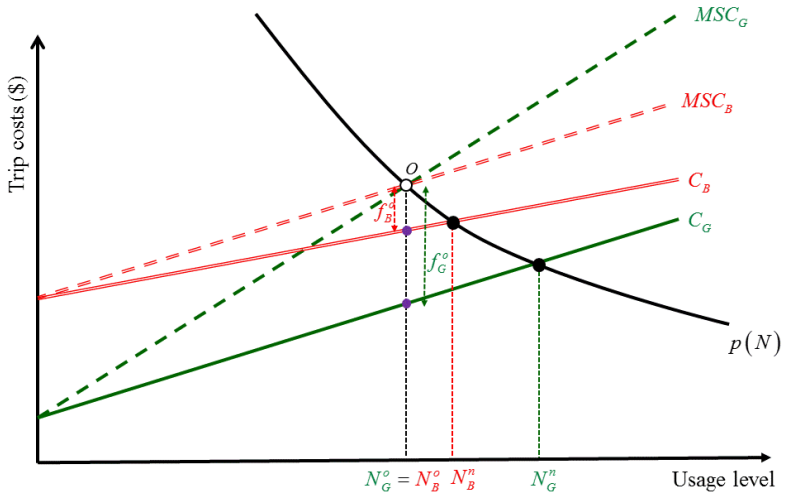

Figure 5: TPS outperforms fee with additive and multiplicative cost shocks

intersect. Consequently, FBO usage is the same in the two states and it can be supported with a TPS by setting $Y^* = N_G^o = N_B^o$. By contrast, a fixed fee cannot support the FBO because the FBO fee is lower in state B than state G. In this example, the net demand curve and marginal external cost curve are positively correlated. Positive correlation works in favour of quantity control — again consistent with eqn. (2) in Weitzman (1974).

5.4 Volatility of tradable permit schemes and fees

Volatility is generally considered to be an undesirable feature of markets. In our setting, volatility can be measured for total usage, the monetary cost of usage, and the full cost of usage. Some comparisons are clear-cut. Total usage varies from state to state when usage is rationed with a fee, but not when it is rationed using a TPS if the permit constraint always binds. In contrast, the monetary cost of gaining access to the facility is constant with a fixed fee, but varies with a TPS because the equilibrium permit price depends on the state.

The relative volatility of the full costs, p^f and p^c, is not as obvious since they vary in both regimes, with the amount of variation depending on the nature of shocks. In the interest of space, attention is limited here to additive shocks. As shown in the Appendix,

$$Var\left(p^f\right) = \frac{d^2 \cdot Var\left(a\right) + b^2 \cdot Var\left(c\right) + 2bd \cdot Cov\left(a, c\right)}{\left(b+d\right)^2},$$

$$Var\left(p^c\right) = Var\left(a\right), \text{ and}$$

$$Var\left(p^c\right) - Var\left(p^f\right) \stackrel{s}{=} \left(b + 2d\right) \cdot Var\left(a\right) - b \cdot Var\left(c\right) - 2d \cdot Cov\left(a, c\right).$$

With a fee, the variance of the full cost increases with the variance of demand, the variance of costs, and the covariance between demand and costs. By contrast, with a TPS the variance of full cost varies one-to-one with the variance of demand, and does not depend on costs. Hence, if only cost is variable the full cost is more volatile with a fee. If only demand is variable, the full cost is more volatile with a TPS. If demand and cost are equally variable, the full cost is more volatile with a TPS unless demand and costs are perfectly and positively correlated. Overall, therefore, the relative volatility of full costs for the two control instruments depends on whether variability originates primarily with demand or primarily with costs.

6 TPS vs. fees: nonlinear functions

The analysis in Section 5 is based on linear demand and cost functions. If either function is nonlinear, Weitzman's rule for ranking the efficiency of prices and quantity controls applies only as a local approximation. Studies of congestible facilities sometimes assume constant-elasticity demand functions which are strictly convex. More important, congestion is often a nonlinear phenomenon. At low usage levels, users may not interfere with each other much, if at all. For example, if the arrival rate of users at a server remains below server capacity, and service times are uniform, no queuing occurs. Crowding at outdoor recreational facilities is typically not considered problematic unless usage exceeds carrying capacity. There is some evidence that visitors do not feel crowded unless the number of contacts exceeds their expectations (Ditton et al., 1983; Michael and Reiling, 1997). There is also evidence that, when crowding does become annoying, disutility grows at an increasing rate with the number of encounters (Boxall et al., 2003). Similarly, airport congestion can be very sensitive to small changes in demand (Jacquillat and Odoni, 2015).

As far as road congestion, traffic engineering studies find that on highways drivers can maintain free-flow speeds until flow reaches a substantial fraction of capacity. Beyond this, speeds can drop rapidly and even unpredictably. Traffic control strategies such as ramp metering and perimeter control are used to avoid breakdown in flow (Menelaou et al., 2017). Following the US Bureau of Public Roads (1964), traffic engineers often specify the relationship between travel time, T, and flow, Q, using a power function of the form $T = T_0\left(1 + d\left(Q/K\right)^{\varepsilon}\right)$, where T_0 is free-flow travel time, K is a measure of road capacity, and $\varepsilon = 4$. If user cost is proportional to travel time, this function can be translated into a cost function of the form $C = c + dN^{\varepsilon}$, which we consider below.

In this section we allow demand and cost functions to be nonlinear. While the analysis is not as straightforward as in Section 5, several insights can be

derived. We begin by showing that Theorems 3 and 4 both generalize. We then identify circumstances where a TPS is superior.

6.1 Potential advantage of a fee

6.1.1 Variable costs

If demand is stationary, a fee outperforms a TPS when certain conditions are met. The key condition is identified in:

Assumption 2 *For any pair of states, one state, ω, is more favorable than another state, $\hat{\omega}$, in the sense that (a) $C_{\omega}(N) \leq C_{\hat{\omega}}(N)$ for all $N \geq 0$, and (b) $C'_{\omega}(N^o_{\omega}) N^o_{\omega} < C'_{\hat{\omega}}(N^o_{\hat{\omega}}) N^o_{\hat{\omega}}$.*

Assumption 2(a) stipulates that at any usage level the user cost is lower in state ω than state $\hat{\omega}$. Assumption 2(b) requires, in addition, that the marginal external cost is lower in state ω than state $\hat{\omega}$ at their respective FBO usage levels. If so, the FBO fee is lower in state ω than state $\hat{\omega}$: $f^o_{\omega}(N^o_{\omega}) < f^o_{\hat{\omega}}(N^o_{\hat{\omega}})$.

Given Assumption 2(a), Assumption 2(b) is plausible insofar as states with high private costs are likely to have high external costs. Nevertheless, if no restrictions are imposed on the demand function, Assumption 2(b) is assured only under fairly restrictive conditions on the cost function. As explained in the Appendix, a necessary condition is that the cost curve be steeper in the less favorable state; i.e. $C'_{\hat{\omega}}(N) > C'_{\omega}(N)$ for all $N > 0$. A sufficient condition to satisfy Assumptions 2(a) and 2(b) is that the cost function has the power form $C_{\omega}(N) = c + d_{\omega} N^{\varepsilon}$, where $c \geq 0$ and $\varepsilon > 0$ are constants and only d_{ω} is state-dependent. If $\varepsilon = 1$, this reduces to the linear case which led to Theorem 3. As noted above, the US Bureau of Public Roads (1964) proposed a power function with $\varepsilon = 4$.

If Assumption 2 holds, states can be ranked in order from least favorable to most favorable. If costs are variable, but demand is stationary, a fee then outperforms a TPS:

Theorem 5 *Let Assumption 2 hold. If only costs are variable, a fee outperforms a TPS.*

Theorem 5 is proved in the Appendix. It generalizes Theorem 3 since it applies regardless of the functional forms of the demand and cost functions as long as they satisfy Assumption 2. Similar to Theorem 4, the proof entails showing that given any permit allocation Y for a TPS, there exists a fixed fee that supports a more efficient usage level than the TPS in every state. The fee drives usage below Y in every state for which optimal usage is below Y, and it supports usage above Y in every state for which optimal usage is above Y.

6.1.2 Variable demand

Theorem 4 established that with linear functions a fee outperforms a TPS if the congestion coefficient d is constant and the permit constraint always binds. This result continues to hold for demand curves of arbitrary shape:

Theorem 6 *Assume that the cost curve is linear, the congestion coefficient d is constant, and the permit constraint always binds. A fee then outperforms a TPS.*

Theorem 6 is proved in Figure 6 using a net benefit curve and MEC curve as in Figure 4, and taking a similar approach to the proof of Theorem 4. The MEC curve has a constant slope of d. The dashed line through points f, c, and h has a slope of $-d$. Dotted line je has the same slope. The optimal permit constraint is Y^*, and a fixed fee is imposed of $\hat{f} = dY^*$. If the net benefit curve happens to cross the MEC curve at point b, both the TPS and the fee support the optimum. If the net benefit curve crosses elsewhere, neither instrument supports the optimum. In Figure 6, the net benefit curve crosses at point c so that optimal usage is $N^o > Y^*$. The deadweight loss from insufficient usage with the TPS is measured by area abc. The deadweight loss from excessive usage with the fee equals area ced. Now

$$abc \underbrace{>}_{(1)} fbc \underbrace{=}_{(2)} chg \underbrace{>}_{(3)} jed \underbrace{>}_{(4)} ced. \tag{40}$$

Inequalities (1) and (4) hold because the net benefit curve is steeper than the MEC curve. Equality (2) is obvious. Finally, inequality (3) applies because triangles chg and jed are equiangular, and side gh is longer than side de. The chain of inequalities in (40) proves that the fee $\hat{f}$ outperforms the TPS. A similar figure can be used to prove that fee $\hat{f}$ is also superior to the TPS when optimal usage is below Y^*. The optimal fee f^*, which generally differs from $\hat{f}$, does not necessarily outperform the TPS in every state, but it does yield (even) higher expected welfare.

In summary, Theorem 6 shows that a fee outperforms a TPS regardless of the shape of the inverse demand curve or how it varies from state to state. The reasoning is the same as for Figure 4 and Laffont's result: the net benefit curve is variable and it is steeper than the MEC curve.

6.2 Potential advantage of a TPS

Theorems 5 and 6 establish fairly general conditions under which a fee outperforms a TPS. Nevertheless, they are limited in scope. Theorem 5 does not allow demand to vary, and it relies on Assumption 2 which does not hold in many instances. It does not hold in Figures 3, 4, or 5, and in Figure 5 the TPS outperforms a fee. Theorem 6 requires the cost function to be linear, and it rules out capacity shocks that affect the congestion coefficient.

We now present an example with variable demand and a fixed, but nonlinear cost function in which a TPS dominates a fee. In the example, shown in Figure 7, the facility has a maximum capacity of K so that the cost curve and MSC curve both become vertical at $N = K$. There are two states of demand, G and B. In both states, the capacity constraint binds in both the unregulated equilibrium and the FBO. To support the FBO without some form of non-price

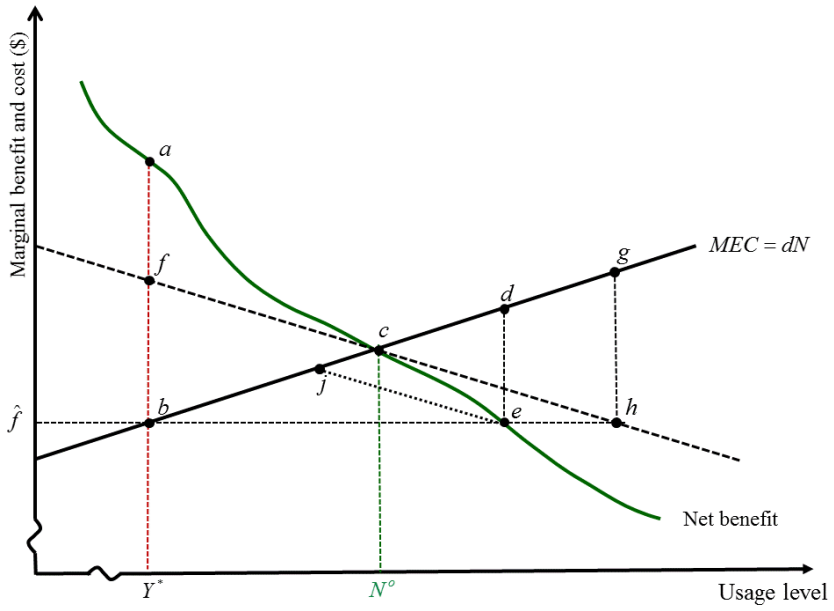

Figure 6: Proof of Theorem 6

rationing such as queuing, a fee must be imposed that is higher in state G than state B. A fixed fee cannot support the optimum in both states. Yet, with a TPS, the FBO can be achieved simply by setting $Y^* = K$.[37] If the planner does not know capacity precisely, it could adopt a margin of safety by setting the TPS at, say, 95% of estimated capacity.[38]

This example shows that a TPS can outdo a fee if the cost function is convex, or becomes sufficiently steep. FBO usage levels then vary little across states. Viewed another way, the MEC curve can be steeper than the net benefit curve. Note that with a linear demand curve, $p = a - bN$, and the power cost function $C = c + dN^{\varepsilon}$, the absolute slope of the net demand curve is $b + d\varepsilon N^{\varepsilon-1}$. The absolute slope of the MEC curve is $d\varepsilon^2 N^{\varepsilon-1}$. The larger is ε, the steeper the MEC curve relative to the net benefit curve.

To explore this reasoning further, we use a numerical example that is descriptive of peak-period automobile commuting. The cost function does not vary, and has the functional form $C(N) = c + dN^{\varepsilon}$. Demand is subject to

[37] This result is consistent with Akamatsu and Wada (2017) who show that a planner who is uncertain about demand can still support the social optimum using permits, but not a fee because the optimal fee depends on demand.

[38] As noted above, traffic engineers often restrict traffic movements to avoid flow breakdown. Hall (2018) argues that the same policy could be adopted with a congestion charge. However, Anderson and Davis (2018) have recently challenged the claim that heavy demand causes flow to break down. They provide empirical evidence that drops in capacity are caused by supply shocks such as road construction, disabled vehicles, and bad weather.

either additive or multiplicative shocks with two states: good days (G) and bad days (B). Parameters with fixed values are $a_B = 40$, $b_B = 0.002$, and $c = 8$. For additive shocks, $a_G = 50$, and for multiplicative shocks, $b_G = 0.0016$. The probability of a good day is set to either 0.2 or 0.8. Parameter ε governing the curvature of the cost function is set to 1, 2, 3, or 4.[39] Parameter d is adjusted to maintain a relatively constant usage level. There are 16 cases in all. For the eight cases with $\varepsilon = 1$ or $\varepsilon = 2$, the fee outperforms the TPS. For the other eight cases with $\varepsilon = 3$ or $\varepsilon = 4$, the TPS outperforms the fee. This provides some support for the conjecture that a TPS has an advantage over a fee when cost functions are sharply curved. de Palma et al. (2018) obtain a similar result in a numerical study of route choice.

In summary, we have shown that a fee outperforms a TPS under Weitzman's assumptions (i.e., with linear demand and cost functions, additive shocks, and a binding permit constraint). A fee is also generally superior with multiplicative shocks. However, a TPS may be superior if the permit constraint does not always bind, if the cost function is strictly convex, or if usage is bounded by a capacity constraint.

7 An adaptive TPS

So far, the comparison of TPS and fees has been limited to a dichotomous choice between basic schemes. The fee is set at a fixed amount per usage that does not depend on either the state or total usage. Similarly, the number of permits allocated each day is independent of the state. Both schemes can be improved at the cost of additional complexity. For example, Weitzman (1978) considered a combination of price and quantity regulation. Kaplow and Shavell (2002) proposed a nonlinear tax scheme for pollution in which the tax rate is chosen to concide with the marginal pollution damage curve. This tax schedule is superior to pure quantity control because quantity control is, in effect, a specific nonlinear tax with no charge for emissions below the target, and an infinite charge above it.

In this section we briefly explore an adaptive version of the basic TPS that effectively combines quantity and price control. The modified TPS is inspired by a pollution control scheme studied by Roberts and Spence (1976).[40] In their scheme, the government first issues tradable licenses. It then imposes a per-unit tax on any firm that emits more than its license holding, and grants a per-unit subsidy to any firm that emits less than its holding. The tax protects firms against very high abatement costs, while the subsidy gives them an incentive to

[39] In traffic engineering studies, parameter ε is usually set between 2 and 5. In our setting, the pertinent relationship is between usage cost and the number of users on a given day or other time interval, and the duration of the usage period may be endogenous. Depending on the structure of trip-timing preferences, the equilibrium cost function can be linear (Arnott et al., 1993), near-linear (de Palma and Marchal, 1999), quadratic (Braid, 1996), or some other function.

[40] Czerny (2008) presents this scheme diagrammatically, but does not examine it analytically or derive its welfare performance relative to a fixed fee or basic TPS.

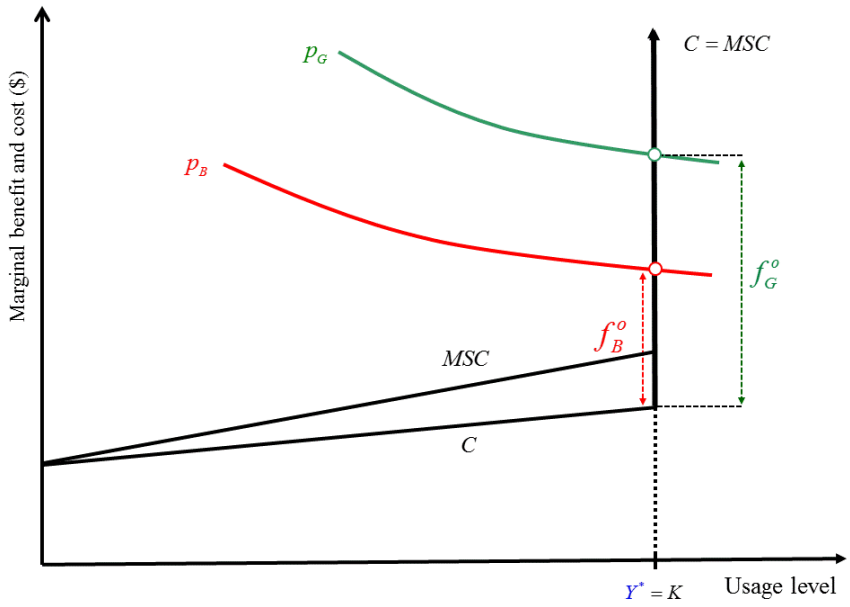

Figure 7: Permit outperforms fee: Example with capacity constraint

abate further if abatement costs turn out to be low.

The adaptive version of the TPS considered here operates similarly. The government issues Y permits, as before. In addition, it offers to sell further permits at a price s, and buy permits at a price r, where $r < s$. Offers to buy and sell act as a collar on the price of permits by limiting it to the range $[r, s]$. Since the values of r and s are fixed, the parameters defining the TPS are independent of the state. However, the TPS is adaptive because users can buy or sell permits once they learn the state.[41] Let q denote the equilibrium price of permits with the collar, and q_0 the price at which permits would trade without the collar. Depending on the state, three outcomes are possible. If $q_0 \in [r, s]$, no government trades occur. If $q_0 < r$, the government buys permits which raises the price to $q = r$. Finally, if $q_0 > s$, the government sells additional permits which drops the price to $q = s$.

The optimal adaptive TPS is derived in the Appendix for the linear model with demand shocks when the composite parameter A is uniformly distributed on the interval $[A_0, A_1]$. The solution is nondegenerate in the sense that, for all distributions with $A_1 > A_0$, the government buys permits when A is close to A_0, sells permits when A is close to A_1, and does not trade when A takes

[41] As Schmalensee and Stavins (2017) explain, the Regional Greenhouse Gas Initiative in the northeastern United States uses an auction that operates in a similar way. When auction prices reach a specified level, additional allowances are sold. There is also a price floor below which allowances are not sold at auction.

intermediate values.

The allocative efficiency of the adaptive TPS can be compared with the efficiencies of the optimal flat fee and basic TPS using the index of relative efficiency

$$e^i = \frac{EW^{i*} - EW^n}{EW^o - EW^n}, \;\; i = f, c, a,$$

where a denotes the adaptive TPS. Index e^i measures the efficiency gain from scheme i as a fraction of the maximum possible gain that can be achieved in moving from the unregulated equilibrium to the social optimum, $EW^o - EW^n$. The indexes work out to

$$e^c = 1 - \frac{Var\,(A)}{\overline{A^2}} \left(\frac{b+d}{d} \right)^2, \tag{41}$$

$$e^f = 1 - \frac{Var\,(A)}{\overline{A^2}}, \tag{42}$$

$$e^a = 1 - \frac{Var\,(A)}{\overline{A^2}} \left(\frac{b+d}{2b+3d} \right)^2. \tag{43}$$

All three schemes are fully efficient when there are no shocks (i.e., when $Var\,(A) = 0$). All fall short of full efficiency when demand or cost is variable. Consistent with Theorem 1, the fee is more efficient than the basic TPS. However, the adaptive TPS is more efficient than the fee. The gap from full efficiency for the adaptive TPS is smaller by a factor $\frac{1-e^a}{1-e^f} = \left(\frac{b+d}{2b+3d} \right)^2$. Depending on the relative size of b and d, this factor ranges from 1/4 down to 1/9.

Figure 8 compares the monetary price with a flat fee, first-best fee, and adaptive TPS for a numerical example with $b = 1$, $d = 1$, $A_0 = 25$, and $A_1 = 50$. With the adaptive TPS, the goverment buys permits when $A \in [25, 35)$, sells additional permits when $A \in (40, 50]$, and is inactive when $A \in [35, 40]$. It is active 80% of the time.[42] The adaptive TPS tracks the FBO fee more closely, on average, than the flat fee. This is consistent with Kaplow and Shavell's (2002) argument that nonlinear tax schemes outperform flat taxes.

8 Extensions

In this section we sketch two ways in which the model can be extended. The first concerns lead times in usage decisions, and the second concerns external costs of usage other than congestion.

[42] As shown in the Appendix, regardless of parameter values both buying and selling are more frequent than not trading.

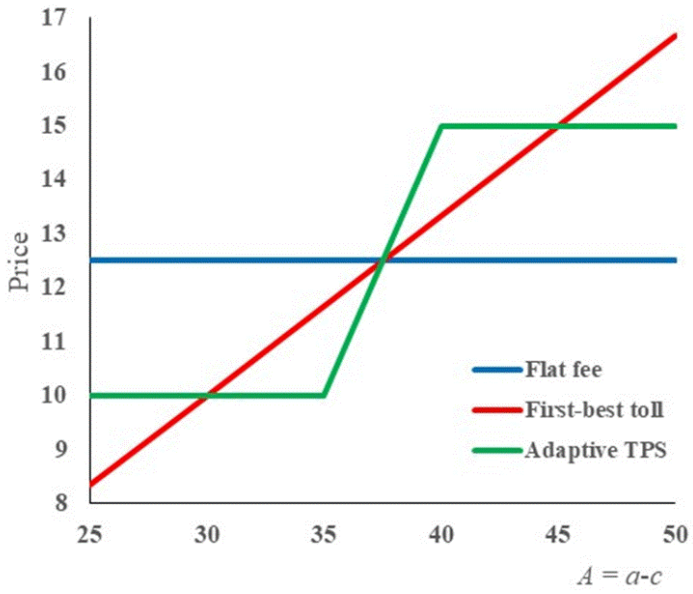

Figure 8: Monetary price of usage with flat fee, first-best fee, and adaptive TPS. $b = 1$, $d = 1$, $A_0 = 25$, $A_1 = 50$.

8.1 Lead times in usage decisions

Usage decisions are often made well before demand and supply conditions are fully known. Commercial airlines typically schedule flights months in advance, and prefer not to cancel them unless circumstances are especially unfavorable. The same is true of railways and trains. Recreationists can plan or book trips weeks or months before they take place, and so on. When usage decisions are made far in advance, they can be conditioned on predictable circumstances such as season and scheduled operating hours, but not unpredictable events such as infrastructure failures or weather at the time of usage.

Lead times in decision-making by users can be accommodated in the model as follows. Let t denote the future time at which usage is considered. For ease of reference, t will be called a day. Agents are assumed to make usage decisions before t, and they neither balk nor make a last-minute decision to use a facility once they learn the actual state at t. Let E_t denote the frequency distribution of days. For each t, there is a probability or frequency distribution of states. Let $E_{\omega|t}$ denote the expectations operator over states conditional on t.

Two cases will be entertained as far as the regulator's decisions. In one, the regulator can adjust the fee and permit allocation according to t, but not ω. For example, flight schedules and hiking permits can be varied by time of year. In this case, the analysis is qualitatively the same for each t as if agents and regulator know only the unconditional distribution of states. It is then straightforward to show that a fee and TPS are equivalent.[43]

In the second case, the regulator cannot condition the fee or permit allocation on either t or ω.[44] Equation (12) determining usage with the fee is then replaced by

$$E_{\omega|t}\left\{p_\omega\left(N_t^f\right)\right\} = E_{\omega|t}\left\{C_\omega\left(N_t^f\right)\right\} + f, \text{ for each } t,$$

and eqn. (13) for the fee is replaced by

$$f^* = \frac{E_t\left\{E_{\omega|t}\left[C'_\omega\left(N_t^f\right)N_t^f\frac{\partial N_t^f}{\partial f}\right]\right\}\Big|_{f^*}}{E_t\left\{\frac{\partial N_t^f}{\partial f}\right\}\Big|_{f^*}}. \tag{44}$$

With the TPS, the equilibrium price of permits depends on t rather than ω. Let Ω_C denote the set of days when the permit constraint binds. Equation (17) for

[43] As noted in Section 2, this is true of Weitzman's model for variations in demand, and Laffont's model for variations in costs.

[44] It may still be possible for a facility manager to adjust operational measures once the state is known. For example, an airport operator can set the maximum number of takeoff and landing operations per hour according to either Visual Meteorological Conditions or Instrument Meteorological Conditions. Similarly, speed limits can be reduced during road work, entrance rates to museum exhibits can be controlled during periods of high demand, and so on. Another possibility is for the regulator to index the fee or permit allowance to some observable measure or signal. For example, carbon emissions can be indexed to national GDP in the form of an emissions intensity cap (Newell and Pizer, 2008). This has the advantage of relaxing the cap when abatement costs are high, and tightening it when the costs are low. The disadvantage is that setting the cap as a function of another, uncertain variable adds noise.

equilibrium usage is replaced by

$$E_{\omega|t}\left\{p_\omega\left(N_t^c\right)\right\} = E_{\omega|t}\left\{C_\omega\left(N_t^c\right)\right\} + q_t, \text{ for each } t,$$

and eqn. (20) is replaced by

$$E_{t\in\Omega_C}\left\{q_t\right\}\big|_{Y^*} = E_{t\in\Omega_C}\left\{E_{\omega|t}\left[C'_\omega\left(N_t^c\right)N_t^c\right]\right\}\big|_{Y^*}. \tag{45}$$

Comparing (44) and (45), it is clear that the fee and TPS are not, in general, equivalent unless usage decisions are the same on every day. Thus, if the regulator cannot adjust instruments to either unpredictable or predictable fluctuations, the two instruments still perform differently under uncertainty even when users cannot adapt to unpredictable fluctuations.

8.2 Additional external costs

Users sometimes create external costs other than congestion. Some external costs are incurred by the population at large such as noise, pollution, greenhouse gas emissions, and damage to flora and fauna. Call them *environmental costs*. Environmental costs generally depend on the amount of usage, and they can also depend on the state.[45] Denote them by $R_\omega(N)$. Other types of external costs are borne by users such as damage to roads and rail track, wear and tear on hiking trails, depletion of fishing stocks, and so on. Call them *damage costs*. Damage costs are a function of cumulative — rather than instantaneous — usage, and they mainly affect future, rather than contemporaneous, users. They can also depend on the state.[46] Denote damage costs incurred per unit of usage by $D_\omega(U)$ where $U \equiv E_\omega\left\{N_\omega\right\}$ is average usage.[47]

Accounting for environmental costs and damage costs, expected welfare is given by an extension of eqn. (11):

$$EW = E\left\{\int_0^{N_\omega} p_\omega(n)\,dn - C_\omega\left(N_\omega\right)N_\omega - D_\omega(U)\,N_\omega - R_\omega\left(N_\omega\right)\right\}. \tag{46}$$

With a fee, the equilibrium usage condition is given by an extension of eqn. (12):

$$p_\omega\left(N_\omega^f\right) = C_\omega\left(N_\omega^f\right) + D_\omega(U) + f. \tag{47}$$

The first-order condition for the optimal fee is

$$E\left\{\left(p_\omega\left(N_\omega^f\right) - C_\omega\left(N_\omega^f\right) - C'_\omega\left(N_\omega^f\right)N_\omega^f - D_\omega(U)\right.\right. \tag{48}$$
$$\left.\left. -D'_\omega(U)\,N_\omega^f - R'_\omega\left(N_\omega\right)\right)\frac{\partial N_\omega^f}{\partial f}\right\}_{f^*} - E\left\{D'_\omega(U)\,N_\omega^f\right\}\cdot E\left\{\frac{\partial N_\omega^f}{\partial f}\right\}\Big|_{f^*} = 0.$$

[45] For example, certain pollutants are more harmful to health during meteorological inversions.

[46] For example, dirt roads and hiking trails are more susceptible to damage after rainstorms than when they are dry. Rail track is more vulnerable to cracking in extreme cold, and more vulnerable to warping in extreme heat.

[47] For a given accounting period, average usage is proportional to cumulative usage. The ecological capacity of a recreational ecosystem is typically a function of total seasonal use, so in this case the relevant accounting period is a year.

Substituting (47) into (48) yields

$$f^* = \frac{E\left\{\left(C'_\omega\left(N^f_\omega\right)N^f_\omega + R'_\omega\left(N_\omega\right)\right)\frac{\partial N^f_\omega}{\partial f}\right\}\Big|_{f^*}}{E\left\{\frac{\partial N^f_\omega}{\partial f}\right\}\Big|_{f^*}} + E\left\{D'_\omega\left(U\right)N^f_\omega\right\}. \tag{49}$$

Eqn. (49) is a generalization of (13).

With a TPS, the equilibrium usage condition is given by an extension of eqn. (17):

$$p_\omega\left(N^c_\omega\right) = C_\omega\left(N^c_\omega\right) + D_\omega\left(U\right) + q_\omega. \tag{50}$$

The first-order condition for the optimal permit allocation is

$$E_{\omega\in\Omega_c}\left\{\begin{array}{c} p_\omega\left(N^c_\omega\right) - C_\omega\left(N^c_\omega\right) - C'_\omega\left(N^c_\omega\right)N^c_\omega - D_\omega\left(U\right) \\ -D'_\omega\left(U\right)N^c_\omega - R'_\omega\left(N_\omega\right)\end{array}\right\}\Big|_{Y^*} = 0,$$

or

$$E_{\omega\in\Omega_c}\left\{q_\omega\right\}|_{Y^*} = E_{\omega\in\Omega_c}\left\{C'_\omega\left(N^c_\omega\right)N^c_\omega + D'_\omega\left(U\right)N^c_\omega + R'_\omega\left(N_\omega\right)\right\}|_{Y^*}. \tag{51}$$

Despite the fact that users do not immediately suffer the damage costs they impose, marginal damage cost appears on the right-hand side of (51) in the same way as the marginal external congestion cost.

Further insight into the implications of damage costs and environmental costs can be gleaned with linear functions. In addition to the demand and cost functions in (21), we adopt

$$D_\omega\left(U\right) = g_\omega U,$$
$$R_\omega\left(N\right) = \left(r_\omega + e_\omega N_\omega\right)N_\omega.$$

Damage costs are assumed to be proportional to usage, whereas total environmental costs are quadratic with $e_\omega > 0$.[48] With additive shocks, expected welfare with the fee is given by a generalization of (28):

$$EW^{f*} = \frac{b+2d+2g}{2\left(b+e+2d+2g\right)^2}\left(\bar{A}-r\right)^2 + \frac{b-2e}{2\left(b+d\right)^2}\sigma^2_A.$$

It the permit constraint always binds, expected welfare with the TPS is given by a generalization of (30):

$$EW^{c*} = \frac{b+2d+2g}{2\left(b+e+2d+2g\right)^2}\left(\bar{A}-r\right)^2.$$

The relative advantage of the fee over the TPS is given by an extension of (31):

$$\Delta = EW^{f*} - EW^{c*} = \frac{b-2e}{2\left(b+d\right)^2}\sigma^2_A.$$

[48] For example, environmental health costs may grow at an increasing rate with the concentration of pollutants. In the case of recreational activities, damage to flora and fauna can mount if hiking paths become so crowded that hikers switch from designated trails to out-of-bound routes (Fleishman et al., 2007). Similarly, at crowded campsites campers may pitch their tents on unprepared sites that are susceptible to damage.

The fee is superior to the TPS if and only if $b > 2e$. Thus, if environmental costs are large enough, the TPS can outperform the fee. To see that this is consistent with Weitzman's rules, note that the inverse net demand curve has a slope $b + d + g$, and the MEC curve has a slope $d + g + 2e$. The MEC curve is steeper than the inverse net demand curve if $2e > b$. When environmental externalities are present as well as congestion, fluctuations in usage are more costly and this favours quantity control — a point that Czerny (2010) mentions. This is an important consideration in large cities, especially in China and India, where the health costs of air pollution are comparable in magnitude to the costs of traffic congestion.

This simple extension of the basic model ignores the possibility that damage can be reduced by maintenance or other conservation activities. If such actions are possible, users will not bear the full costs of damage and the feedback effect of damage in limiting usage will be weakened. The portion of damage costs that users do not incur will enter the social calculus in the same way as environmental costs, and further strengthen the relative performance of a TPS.

9 Conclusions

Tradable Permit Schemes (TPS) have been implemented at the continental, national, and regional level to control carbon emissions and other pollutants. Slot trading and slot auctions have been used to control usage of airports. Academic interest is now growing in the potential use of TPS to regulate road transport. Advances in information and communications technology have made the use of TPS conceivable for roads and other congestible facilities such as recreational areas. TPS also have an advantage in public acceptability over tolls and other user fees since permits can be distributed free so that users in aggregate do not incur an out-of-pocket cost. TPS thus offer a plausible alternative to fees as a means of regulating access to facilities that are prone to overusage.

It is well known that TPS and fees can both support optimal usage of a congestible facility in a stationary environment. What remains relatively unexplored is how the two instruments compare when demand and costs fluctuate, and both the quantity of permits and the fee are constrained to be the same in all states. We find that, in general, a TPS is relatively efficient if optimal usage levels are similar across states. Analogously, a flat congestion fee achieves high efficiency if the first-best fee varies little over states. When usage costs are variable, a fee outperforms a TPS if the optimal congestion fee is higher in states where the usage cost function is higher. A fee also tends to outperform a TPS when the demand and cost functions are linear and do not vary too much. However, under several other circumstances a TPS can outperform a fee. A TPS tends to be superior if the cost function is strongly convex or if usage is bounded by a capacity constraint. A TPS has an edge in flexibility if demand or costs fluctuate sufficiently that the permit constraint does not always bind. And a TPS is advantageous if externalities such as pollution are present that, unlike congestion, do not give agents direct feedback on the socially efficient

level of usage.

Further analysis with nonlinear functions, empirically-based probability distributions of states, and facility–specific characteristics is needed to assess the robustness of these findings. For facilities such as roads that are generally accessible at all times, usage tends to spread out by time of day as demand increases, and equilibrium costs rise more smoothly than for facilities with restricted operating hours. Similarly, for facilities with multiple sites or routes, usage can spread out over space. Downhill skiing areas present a more complicated case because congestion can occur on the slopes, as queues for ski lifts, on connecting trails, and at the lodge (Barro and Romer, 1987).

The model can be extended in various ways. It can encompass other choice dimensions besides the amount of usage including time of use, travel mode for transportation, hiking trail, or visit duration for recreation. Nonatomistic users can also be considered such as major airlines that individually account for a substantial fraction of total airport traffic. As Brueckner (2009) shows, with nonatomistic users congestion fees and TPS are no longer equivalent even without uncertainty.[49]

In the model the regulator cannot condition either the number of permits or the fee on the state, but the regulator does know the probability distribution of states. In practice, this may not be the case. The frequency of floods, windstorms, and other severe natural events is evolving as the climate changes. Nature areas and other ecosystems may be susceptible to catastrophic or irreversible effects. Human-caused shocks such as transit strikes and terrorist attacks are also hard to quantify. The performance of quantity controls and price controls may differ in the face of these uncertainties, and it may be wise to adopt policies that are robust to the worst circumstances.

Yet another possibility is to consider a system of multiple permits in which separate permit constraints are imposed on each facility within a network or group of facilities. Several questions arise in such a setting. Would a multiple permit scheme be welfare-superior to a single integrated scheme? If so, would the advantage be large enough to outweigh the greater administration and compliance costs? What happens if facilities are controlled by different entities? Czerny and Lang (2019) consider two airports with interconnecting flights and local objective functions that independently choose between setting fares and imposing slot controls to constrain traffic volumes. They show that, in general, independent decision-making does not yield a first-best outcome. Each airport ignores the effects of its decision on welfare at the other airport.

We conclude with a few comments on the potential acceptability and equity advantages of TPS relative to tolls for road travel. Since permits can be distributed free, drivers do not have to pay a charge to a government or road administrator. This avoids objections, often raised against tolls, that permits serve as a cash cow, or constitute double taxation. It also addresses equity con-

[49] Nonatomistic or "large" users have an incentive to internalize their self-imposed congestion when they can affect the total amount of usage. Discriminatory charges based on user size are then necessary to support efficient usage (Brueckner, 2002). A TPS is free of this complication, although inefficiencies may arise if large users exercise their market power in trading permits.

cerns that the rich can buy time. Laboratory experiments by Exley and Kessler (2019) indicate that people care more about inequity in time than inequity in money. This may help to explain aversion to policies that allow users to bypass queues on roads, as well as other facilities such as airports, amusement parks, and hospitals. Controlling access using permits rather than tolls is likely to be less controversial.

A potential weakness of TPS is that there is no obviously acceptable way to distribute permits.[50] Our model features a single facility, and potential users are identical other than for their willingness to pay. In reality, travel takes place on extensive road networks at different times of day. Individuals differ in numerous ways: residential and workplace locations, income, opportunity cost of travel time, flexibility in when to travel, frequency of travel, and so on. Depending on the scope of a TPS, only a small fraction of agents may need permits. Allocating them to every resident of a large area would leave few permits in the hands of those who need them. Yet, targeting permits to these individuals may be viewed as inequitable. It also risks distorting behavior in undesirable ways (e.g., encouraging vehicle ownership, changing place of residence or work, altering route, etc.).

Similar issues arise with credit-based congestion pricing (CBCP). CBCP involves paying tolls rather than using permits, and individuals are given money endowments (i.e., credit) to offset the cost of paying tolls. Yet, despite these differences, the distribution of credit raises similar problems to the distribution of permits. Kockelman and Kalmanje (2005) describe how CBCP would work, and summarize public attitudes towards it. Gulipalli et al. (2008) survey expert opinions. Transport economists raised concerns about the distribution of credit. They considered it unfair for everyone with a driver's license to receive credit, whether or not they actually pay tolls. Some transport economists opposed basing credit allocations on income. Doing so might be administratively burdensome, and would make CBCP like a welfare program. Economists generally favor addressing distributional goals using income taxes, rather than distorting prices for goods and services. Allocating credit to vehicle owners, on the other hand, would tend to be regressive. Gulipalli et al. note that basing credit on residential location might be advantageous if tolling is widespread, while basing it on distance driven might be better if tolling is limited to highways. They provisionally settle on basing it on vehicle registration as the best of imperfect alternatives, noting that owners of more than one vehicle should receive only one credit allotment.

In an application of CBCP to Austin, Texas, Kalmanje and Kockelman (2004) assume that credit is distributed to all residents with a valid driver's license. In another application to Dallas-Fort Worth, Gulipalli and Kockelman (2008) consider several allocation mechanisms including to all registered vehicle owners, and restricting credit to commuters who use freeways. Yet another possibility would be to favor residents who lack good access to public transit. In summary, deciding how to allocate credit or permits is a difficult question

[50] We are grateful to an anonymous referee for raising this point.

that requires further research.

10 Acknowledgments

We are grateful to the editor and two anonymous referees for helpful comments and suggestions. We are also indebted to Moshe Ben-Akiva, Vivek Ghosal, José Holguín-Veras, Lucas Javaudin, Juan Pablo Montero, Hugo Silva, and seminar participants at Pontificia Universidad Católica de Chile, Rensselaer Polytechnic Institute, Ecole Normale Supérieure de Cachan, Laval University, VEDECOM (France), The Hong Kong Polytechnic University, and UBC. We also thank participants at the 2017 Transportation Research Forum conference and our discussant Ken Small, participants at the 2017 North American Meetings of the Regional Science Association International and our discussant Stephen McCarthy, and participants at the 2019 annual conference of the Western Economic Association International and our discussant Jonathan Hall. Finally, special thanks are due to Achim Czerny who brought to our attention important references and offered insightful feedback. The usual disclaimer applies.

Financial support from the Social Sciences and Humanities Research Council of Canada (Grant 435-2014-2050) is gratefully acknowledged.

11 References

Adar, Z. and J. M. Griffin (1976). Uncertainty and the choice of pollution control instruments. Journal of Environmental Economics and Management 3(3), 178-188.

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.

American Recreation Coalition (2004). Congress Replaces National Recreation Fee Demonstration Program (http://www.funoutdoors.com /node/view/1208, last accessed August 18, 2018)

Anderson A., and W. Freimund (2004). Multiple dimensions of active opposition to the Recreational Fee Demonstration Program. Journal of Park and Recreation Administration 22(2), 44-64.

Anderson, M. L. and L. W. Davis (2018). Does hypercongestion exist? New evidence suggests not. National Bureau of Economic Research Working Paper No. w24469.

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

Ball, M., C. Barnhart, M. Dresner, M. Hansen, K. Neels, A. Odoni, E. Peterson, L. Sherry, A. Trani and B. Zou (2010). Total Delay Impact Study. Technical report, National Center of Excellence for Aviation Operations Research (http://www.isr.umd.edu/NEXTOR/pubs/TDI_Report_Final_10_18_10_V3.pdf,

last accessed August 18, 2018)

Ball, M., G. Donohue and K. Hoffman (2006). Combinatorial Auctions. MIT Press. Chap: Auctions for the Safe, Efficient, and Equitable Allocation of Airspace System Resources, pp. 507–538.

Barro, R. and P.M. Romer (1987). Ski-lift pricing, with applications to labor and other markets. American Economic Review 77(5), 875-890.

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.

Boxall, P., K. Rollins and J. Englin (2003). Heterogeneous preferences for congestion during a wilderness experience. Resource and Energy Economics 25, 177-195.

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

Brands, D., E.T. Verhoef, J. Knockaert and P. Koster (2019). Tradable permits to manage urban mobility: Market design and experimental implementation. January 27, Tinbergen Institute Discussion Paper 2019-007/VIII (https://ssrn.com/abstract=3323642 or http://dx.doi.org/10.2139/ssrn.3323642)

Brueckner, J.K. (2002). Airport congestion when carriers have market power. American Economic Review 92(5), 1357-1375.

Brueckner, J.K. (2009). Price vs. quantity-based approaches to airport congestion management. Journal of Public Economics 93(5–6), 681-690.

Carlin, A. and P. Park (1970). Marginal cost pricing of airport runway capacity. American Economic Review 60(3), 310–319.

Cicchetti, C.J. and V.K. Smith (1973). Congestion, quality deterioration, and optimal use: Wilderness recreation in the Spanish Peaks primitive area. Social Science Research 2(1), 15-30.

Commission of the European Communities (2008). On the Application of Regulation (EEC) No. 95/93 on Common Rules for the Allocation of Slots at Community Airports, as amended. European Commission, Brussels (http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52008DC0227, last accessed August 18, 2018)

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

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

Czerny, A.I. and H. Lang (2019). A pricing versus slots game in airport networks. Transportation Research Part B: Methodological 125, 151-174.

De Borger, B. and S. Proost (2013). Traffic externalities in cities: The economics of speed bumps, low emission zones and city bypasses. Journal of Urban Economics 76, 53-70.

de Palma, A. and R. Lindsey (1998). Information 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 19(6), 1377-1399.

de Palma, A. and F. Marchal (1999). Analysis of travel cost components using large-scale, dynamic traffic models. Transportation Research Record 1676, 177-183.

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.

Ditton, R.B., A.J. Fedler and A.R. Graefe (1983). Factors contributing to perceptions of recreational crowding. Leisure Sciences 5(4), 273-288.

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.

Ecola, L. and T. Light (2009). Equity and congestion pricing: A review of the evidence. Technical Report, Rand Transportation, Space, and Technology. (http://www.rand.org/pubs/technical_reports/TR680/, last accessed August 18, 2018)

Espey, M. (2006). Implementation of recreation fees by the U.S. Forest Service: 1996-2002. Journal of Park and Recreation Administration 24(2), Summer, 87-101.

Exley, C.L. and J.B. Kessler (2019). Equity concerns are narrowly framed. NBER Working Paper No. 25326
(https://www.nber.org/papers/w25326)

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

Fishelson, G. (1976). Emission control policies under uncertainty. Journal of Environmental Economics and Management 3, 189-197.

Fishelson G. and F. Flatters (1975). The non equivalence of tariffs and quotas under uncertainty. Journal of International Economics 5(4), 385-393.

Fleishman, L., E. Feitelson and I. Salomon (2007). Behavioral adaptations to crowding disturbance: Evidence from nature reserves in Israel. Leisure Sciences 29, 37–52.

Fukui, H. (2010). An empirical analysis of airport slot trading in the United States. Transportation Research Part B 44(3), 330-357.

Gillen, D., A. Jacquillat and A. R. Odoni (2016). Airport demand management: The operations research and economics perspectives and potential synergies. Transportation Research Part A 94, 495-513.

Goodkind, A.L. and J.S. Coggins (2015). The Weitzman price corner. Journal of Environmental Economics and Management 73, 1-12.

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

Gu, Y., E. Deakin and Y. Long (2017). The effects of driving restrictions on travel behavior evidence from Beijing. Journal of Urban Economics 102, 106-122.

Gulipalli, P.K., S. Kalmanje and K.M. Kockelman (2008). Credit-based congestion pricing: Expert expectations and guidelines for application. Journal

of the Transportation Research Forum 47(2), 5-19.

Gulipalli, P.K. and K.M. Kockelman (2008). Credit-based congestion pricing: A Dallas-Fort Worth application. Transport Policy 15, 23-32.

Hall, J.D. (2018). Pareto improvements from Lexus Lanes: The effects of pricing a portion of the lanes on congested highways. Journal of Public Economics 158, 113-125.

He, F., Y. F. Yin, N. Shirmohammadi and Y. Nie (2013). Tradable credit schemes on networks with mixed equilibrium behaviors. Transportation Research Part B: Methodological 57, 47–65.

Hoel, M. and L. Karp (2001). Taxes and quotas for a stock pollutant with multiplicative uncertainty. Journal of Public Economics 82, 91-114.

Inrix (2016). Traffic congestion cost UK motorists more than £30 billion in 2016 (http://inrix.com/press-releases/traffic-congestion-cost-uk-motorists-more-than-30-billion-in-2016/, last accessed August 18, 2018).

Jacquillat, A. and A. Odoni (2015). Endogenous control of arrival and departure service rates in dynamic and stochastic queuing models with application at JFK and EWR. Transportation Research Part E: Logistics and Transportation Review 73(1), 133–151.

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.

Kalmanje, S. and K.M. Kockelman (2004). Credit-based congestion pricing: travel, land value, and welfare impacts. Transportation Research Record 1864, 45-53.

Kaplow, L. and S. Shavell (2002). On the superiority of corrective taxes to quantity regulation. American Law and Economics Review 4, Spring, 1-17.

Kockelman, K.M. and S. Kalmanje (2005). Credit-based congestion pricing: A policy proposal and the public's response. Transportation Research Part A: Policy and Practice 39(7–9), 671–690.

Laffont, J.J. (1977). More on prices vs. quantities. Review of Economic Studies 44, 177–182.

Levine, M.E. (2008). Airport congestion: When theory meets reality. Yale Journal on Regulation 26(1), 37-88.

Levinson, D.M. and A. Odlyzko (2008). Too expensive to meter: The influence of transaction costs in transportation and communication. Philosophical Transactions of the Royal Society A 366(1872), 2033–2046.

Manning, R. (1999). Studies in Outdoor Recreation, 2nd edition, Corvallis: Oregon State University Press.

Menelaou, C., S. Timotheou, P. Kolios and C.G. Panayiotou (2017). Improved road usage through congestion-free route reservations. Transportation Research Record: Journal of the Transportation Research Board 2621, 71–80.

Merrill Brown, L., A. Hanafi and A. Petsonk (2012), The EU Emissions Trading System: Results and Lessons Learned, Executive Summary, Environmental Defense Fund (https://www.edf.org/sites/default/files/EU_ETS_Lessons_Learned_Executive_Summary_EDF.pdf, accessed August 7, 2018)

Michael, J.A. and S.D. Reiling (1997). The role of expectations and heterogeneous preferences for congestion in the valuation of recreation benefits. Agricultural and Resource Economics Review 26(2), 166-273.

More, T., D. Dustin and R. Knopf (1996). Behavioral consequences of campground user fees. Journal of Park and Recreation Administration 14(1), 81-93.

Newell, R. G. and W. A. Pizer (2008). Indexed regulation. Journal of Environmental Economics and Management 56(3), 221-233.

Noordegraaf, D.V., B. Heijligers, O. van de Riet and B. van Wee (2009). Technology options for distance-based road user charging schemes. Paper Presented at the 88th Annual Meeting of the Transportation Research Board, Washington, DC. Conference CD Paper No. 09-2477. (https://trid.trb.org/view/881759, last accessed August 18, 2018).

Padmanabhan, V., I. Tsetlin and T. Van Zandt (2010). Setting price or quantity: Depends on what the seller is more uncertain about. Quantitative Marketing and Economics 8, 35–60.

Pigou, A.C. (1920). The Economics of Welfare, London: Macmillan.

Roberts, M.J. and M. Spence (1976). Effluent charges and licenses under uncertainty. The Journal of Public Economies 5(3-4), 193-208.

Rose-Ackerman, S. (1973). Effluent charges: A critique. Canadian Journal of Economics 6, 512-527.

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, 561-583.

Schmalensee, R. and R. N. Stavins (2017). Lessons learned from three decades of experience with cap and trade. Review of Environmental Economics and Policy 11(1), 59-79.

Schrank, D., B. Eisele, T. Lomax and J. Bak (2017). 2015 Urban Mobility Scorecard. Technical report, Texas A&M Transportation Institute and INRIX, August.

Scrogin, D. (2005). Lottery-rationed public access under alternative tariff arrangements: changes in quality, quantity, and expected utility. Journal of Environmental Economics and Management 50(1), 189

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, 27–36.

Smith, V.K. and J.V. Krutilla (1974). A simulation model for the management of low density recreational areas. Journal of Environmental Economics and Management 1(3), 187-201.

Stavins, R.N. (1997). Correlated uncertainty and policy instrument choice. Journal of Environmental Economics and Management 30, 218-232.

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

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

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.

Victoria Transport Policy Institute (2014). Vehicle Restrictions (http://www.vtpi.org/tdm/tdm33.htm, last accessed August 18, 2018)

Watson, A.E. and G. Herath (1999). Research implications of the theme issues 'recreation fees and pricing issues in the public sector' (Journal of Park and Recreation Administration) and 'societal response to recreation fees on public lands' (Journal of Leisure Research)". Journal of Leisure Research 31(3), 325-334.

Watson, W. and R. Ridker (1984). Losses from effluent taxes and quotas under uncertainty. Journal of Environmental Economics and Management 11, 310–326.

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

Weitzman, M.L. (1978). Optimal rewards for economic regulation. American Economic Review 68(4), 683-691.

Weyl, E.G. (2012). Prices v. quantities: Teaching note. University of Chicago. September (https://www.dropbox.com/s/t8fqd5hhi4rplpr/note_pvq.pdf?dl=0, last accessed August 18, 2018)

Yang, H. and X. Wang (2011). Managing network mobility with tradable credits. Transportation Research Part B 45, 580–594.

Yohe, G. W. (1978). Towards a general comparison of price controls and quantity controls under uncertainty. Review of Economic Studies 45, 229-238.

Yoshimura, Y., A. Krebs and C. Ratti (2017). Noninvasive Bluetooth monitoring of visitors' length of stay at the Louvre. Pervasive Computing, IEEE CS, 26-34.

Yoshimura, Y., S. Sobolevsky, C. Ratti and R. Sinatra (2014). An analysis of visitors' behavior in The Louvre Museum: a study using Bluetooth data. Environment and Planning B Planning and Design 41(6), 1113-1131.

12 Notational glossary

Latin characters

a, b, c, d, e, g, r : parameters of cost functions
$C(N)$: user cost function
$C_\omega(.)$: user cost function in state ω
$D_\omega(.)$: damage function in state ω
E_ω : expectations operator
f : usage fee
MEC : marginal external cost
MSC : marginal social cost
N : total usage
$p(N)$: willingness to pay for usage
$\wp_\omega$: probability of state ω, $\omega \in \Omega$
q : price of one permit
$R_\omega(.)$: environmental cost function in state ω
t : date of usage
U : average usage
W : social surplus or welfare
Y : total number of permits distributed

Regimes (denoted by superscripts)

c : permit (or credit)
f : fee
i : index of regimes
n : unregulated
o : first-best optimum

Greek characters

ω : a particular state, with $\omega \in \Omega$
Ω : set of possible states
Ω_C : set of states in which permit constraint binds
Ω_N : set of states in which permit constraint does not bind

A Appendix (for online publication)

A.1 Bounds on the optimal fixed fee (Section 4.1)

Usage with the FBO fee in state ω, N_ω^o, is defined implicitly by the condition

$$p_\omega\left(N_\omega^o\right) = C_\omega\left(N_\omega^o\right) + f_\omega^o.$$

Usage with the fixed fee in state ω, N_ω^f, is defined implicitly by the condition

$$p_\omega\left(N_\omega^f\right) = C_\omega\left(N_\omega^f\right) + f.$$

If $f > \bar{f}_\omega^o$ then $N_\omega^f < N_\omega^o$ for all $\omega \in \Omega$. Welfare can be improved by marginally reducing f in order to increase N_ω^f closer to N_ω^o in every state. Similarly, if $f < \underline{f}_\omega^o$ then $N_\omega^f > N_\omega^o$ for all $\omega \in \Omega$. Welfare can be improved by marginally increasing f in order to reduce N_ω^f closer to N_ω^o in every state.

A.2 Example of nonunique optimal fixed fee (Section 4.1)

The example features linear functions as in Section 5. The cost function, $C(N) = dN$, is stationary. There are two states of demand, B and G, which occur with probabilities $\wp$ and $1-\wp$, respectively. In state B, the inverse demand curve is $p_B(N) = a_B - bN$. In state G, the inverse demand curve is[51]

$$p_G(N) = \begin{cases} a_G - bN & \text{for } N \leq (a_G - \Gamma)/(b+d) \\ 0 & \text{for } N > (a_G - \Gamma)/(b+d) \end{cases}.$$

Parameter values are such that $a_G > a_B$, and $\Gamma \in (f_B^o, f_G^o)$ where f_ω^o is the first-best fee for state ω, given in eqn. (25). The example is constructed so that two distinct fees can yield the same expected welfare. One fee is $f = f_B^o$ which supports optimal usage in state B, N_B^o, without affecting usage in state G. The other fee satisfies $f \in (\Gamma, f_G^o)$. This larger fee reduces usage in state G towards N_G^o while driving usage in state B below N_B^o. It is straightforward to show that the two fees yield the same expected welfare if parameter values satisfy:

$$\left(\wp\left(a_G - a_B\right)^2 - a_G^2\right) d^2 + 2d\left(b + 2d\right) a_G \Gamma - \left(b + 2d\right)^2 \Gamma^2 = 0.$$

One such instance is $a_G = 20$, $a_B = 10$, $b = 1$, $d = 1$, $\Gamma = 5$, and $\wp = 0.25$. The low-fee solution is

$$f = f_B^o = 3.\dot{3},\ N_B = N_B^o = 3.\dot{3},\ N_G = 7.5.$$

The deadweight loss from overusage in state G is shown by area def in Figure 9. The high-fee solution is

$$f = 5.8\dot{3},\ N_B = 2.08\dot{3},\ N_G = 7.08\dot{3}.$$

[51] Contrary to what is assumed in Section 3, the inverse demand curve in state G is not differentiable. It can be closely approximated by a smooth function that still illustrates nonuniqueness.

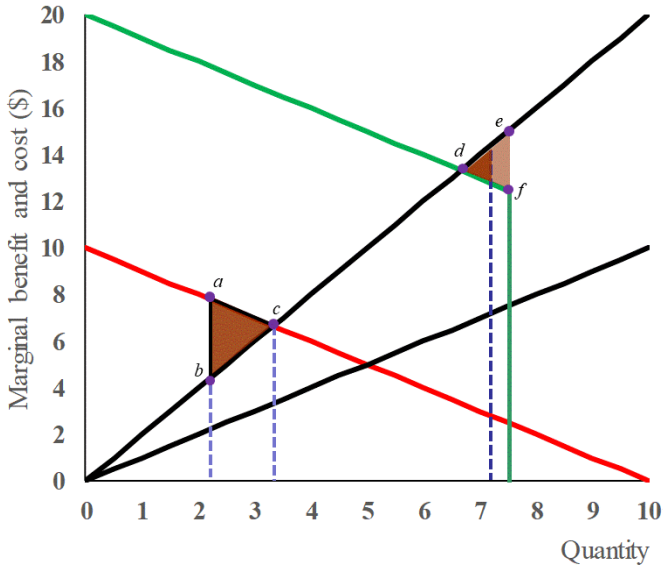

Figure 9: Example with nonunique flat fee

The deadweight loss from underusage in state B is shown by the dark area abc, and the deadweight loss from overusage in state G is shown by the small dark area within area def. The high fee balances the two areas because state G is three times as likely as state B. Expected welfare for both fees is $EW \cong 53.385$.

A.3 Positive usage with the optimal fixed fee (Section 5.1)

Unregulated equilibrium usage is given by eqn. (23): $N_\omega^n = A_\omega / (b+d)$. The permit constraint is (29): $Y^* = \bar{A} / (b+2d)$. The constraint binds if

$$\frac{A_\omega}{\bar{A}} > \frac{b+d}{b+2d}. \tag{A.1}$$

The optimal fixed fee is given by eqn. (27): $f^* = d\bar{A} / (b+2d)$. Usage is $N_\omega = (A_\omega - f^*) / (b+d)$, which is positive if

$$\frac{A_\omega}{\bar{A}} \geq \frac{d}{b+2d}. \tag{A.2}$$

Condition (A.2) is satisfied strictly if (A.1) holds.

A.4 Proof of Theorem 2 (Section 5.1)

Consider an example with two states, G and B. Expected welfare for the two states in the unregulated and FBO regimes is derived from eqn. (24):

$$EW^n = \frac{b}{2(b+d)^2}\left(\wp A_B^2 + (1-\wp)A_G^2\right),$$

$$EW^o = \frac{1}{2(b+2d)}\left(\wp A_B^2 + (1-\wp)A_G^2\right).$$

Expected welfare with the fee is given by eqn. (28):

$$EW^{f*} = \frac{\left(\wp A_B + (1-\wp)A_G\right)^2}{2(b+2d)} + \frac{b}{2(b+d)^2}\wp(1-\wp)(A_G - A_B)^2. \quad \text{(A.3)}$$

This formula is applicable only if demand in state B is positive, which requires

$$\frac{A_B}{A_G} > \frac{(1-\wp)d}{b+(2-\wp)d}. \quad \text{(A.4)}$$

Expected welfare with a TPS that binds in both states (and achieves at least a local optimum) is given by eqn. (30):

$$EW_{BG}^c = \frac{1}{2(b+2d)}\left(\wp A_B + (1-\wp)A_G\right)^2.$$

The TPS binds in both states if $N_B^n \geq Y^*$. Given eqns. (23) and (29) this requires

$$\frac{A_B}{A_G} \geq \frac{(1-\wp)(b+d)}{b+2d-\wp(b+d)}.$$

Finally, expected welfare with the TPS when it binds only in state G with $Y = N_G^o$ is derived from eqn. (24):

$$EW_G^c = \frac{b}{2(b+d)^2}\wp A_B^2 + \frac{1}{2(b+2d)}(1-\wp)A_G^2. \quad \text{(A.5)}$$

The TPS binds only in state G if $N_B^n < N_G^o$. Given eqns. (23), this requires

$$\frac{A_B}{A_G} < \frac{b+d}{b+2d}. \quad \text{(A.6)}$$

Choosing the TPS to bind only in state G is welfare-superior if $EW_G^c > EW_{BG}^c$, or

$$\frac{A_B}{A_G} < \frac{\sqrt{1-\wp}}{\sqrt{1-\wp} + d/(b+d)}. \quad \text{(A.7)}$$

The RHS of inequality (A.7) exceeds the RHS of inequality (A.4). A range of values of A_B/A_G therefore exists such that both constraints are satisfied.

The TPS is welfare-superior to the fee if $EW_G^c > EW^{f*}$. Using eqns. (A.3) and (A.5) this implies

$$\frac{A_B}{A_G} < \frac{\sqrt{1-\wp} - (1-\wp)}{\wp}. \tag{A.8}$$

Given Theorem 1, condition (A.7) must hold when (A.8) is satisfied. The RHS of inequality (A.8) exceeds the RHS of inequality (A.4). Hence the nonnegativity condition on demand with the fee can be satisfied when condition (A.8) holds. Note, finally, that condition (A.6) is satisfied when (A.8) is satisfied. If condition (A.4) fails, then both the fee and the TPS are optimized for state G and support the same usage level. ■

A.5 Proof of Theorem 3 (Section 5.2)

Both the cost function, $C_\omega(N) = c + d_\omega N$, and the FBO fee, $f_\omega^o(N_\omega^o) = C_\omega'(N_\omega^o) N_\omega^o = (a-c) d_\omega / (b + 2d_\omega)$, are increasing function of d_ω. States with lower costs (i.e., lower d_ω) thus also have lower marginal external costs and FBO fees. Assumption 2 is therefore satisfied and Theorem 5 applies. ■

A.6 Proof of Theorem 4 (Section 5.2)

In Theorem 4, parameters a, b, and c can vary with the state, but d is constant. Consider any state ω. To economize on notation, the dependence of variables on ω is suppressed. Welfare is given by eqn. (22). If the permit constraint binds, then $N = Y$ and welfare with the TPS is

$$W^c = AY - \frac{b+2d}{2} N^2. \tag{A.9}$$

Suppose the fixed fee is set to $\hat{f} = dY$. Usage is then $N = (A - dY) / (b + d)$, and welfare with the fee is

$$\begin{aligned} W^f &= A\left(\frac{A-dY}{b+d}\right) - \frac{b+2d}{2}\left(\frac{A-dY}{b+d}\right)^2 \\ &= \frac{bA^2}{2(b+d)^2} + \frac{Ad^2}{(b+d)^2} Y - \frac{d^2(b+2d)}{2(b+d)^2} Y^2. \end{aligned} \tag{A.10}$$

Given (A.9) and (A.10), the difference in welfare is

$$W^f - W^c = \frac{b}{2(b+d)^2} (A - (b+2d) Y)^2 \geq 0.$$

Welfare is strictly higher with the fee except in states for which $A_\omega - (b_\omega + 2d_\omega) Y = 0$ (i.e., for which $Y = N_\omega^o$). Hence, unless optimal usage is the same in all states the fee strictly outperforms the TPS. ■

The proof of Theorem 4 uses the fee $\hat{f} = dY$. The optimal fee, f^*, can be derived using eqn. (14); it works out to

$$f^* = d\frac{E\left\{\frac{A_\omega}{(d+b_\omega)^2}\right\}}{E\left\{\frac{1}{d+b_\omega}\right\} + dE\left\{\frac{1}{(d+b_\omega)^2}\right\}}.$$

The optimal permit allocation can be derived using eqns. (17) and (19):

$$Y^* = \frac{\bar{A}}{E\left\{b_\omega\right\} + 2d}.$$

In general, $f^* \neq dY^*$. Theorem 4 can also be proved by comparing expected welfare with the fee f^*, and expected welfare with the permit allocation set to Y^*.

A.7 Example with multiplicative demand shocks in which TPS outperforms fee (Section 5.2)

Consider an example in which only the slope of the demand curve is variable. Parameter b has a two-point distribution: $b = b_B$ with probability $\wp$, and $b = b_G$ with probability $1 - \wp$ where $b_B > b_G > 0$. Demand is higher in state G than state B.

Suppose the permit allocation is set to support the FBO in state G. The permit constraint does not bind in state B if $b_B > b_G + d$. Using eqns. (34), welfare with the TPS is

$$EW^c = A^2\left(\frac{\wp b_B}{2\left(b_B+d\right)^2} + \frac{1-\wp}{2\left(b_G+2d\right)}\right).$$

Using eqn. (35), welfare with the fee is

$$EW^{f*} = \frac{A^2}{2}\frac{\left(\frac{\wp}{b_B+d} + \frac{1-\wp}{b_G+d}\right)^2}{\frac{\wp(b_B+2d)}{(b_B+d)^2} + \frac{(1-\wp)(b_G+2d)}{(b_G+d)^2}}.$$

Suppose $\wp = 0.5$, $d = 0.5$, and $b_G = 1$. Then $EW^{f*} < EW^c$ if $b_B > 3.31$. The TPS dominates when demand is so low on bad days that it is optimal to ignore bad days and support the FBO on good days.

A.8 Volatility of the full price of usage (Section 5.4)

A.8.1 Fee regime

The full price of usage is $p^f = a - bN^f$ with $N^f = (A - f)/(b + d)$. Substituting for f using eqn. (27) gives $p^f = \left(da + bc + bd\bar{A}/(b + 2d)\right)/(b + d)$, which has a variance

$$Var\left(p^f\right) = \frac{d^2 \cdot Var\left(a\right) + b^2 \cdot Var\left(c\right) + 2bd \cdot Cov\left(a, c\right)}{\left(b+d\right)^2}. \tag{A.11}$$

A.8.2 TPS regime when permit constraint always binds

The full price of usage is $p^c = a - bY$ with $Y = \bar{A}/(b+2d)$ from eqn. (29). Hence $p^c = a - b\bar{A}/(b+2d)$ which has a variance

$$Var(p^c) = Var(a). \tag{A.12}$$

Given (A.11) and (A.12),

$$Var(p^c) - Var\left(p^f\right) \stackrel{s}{=} (b+2d) \cdot Var(a) - b \cdot Var(c) - 2d \cdot Cov(a,c).$$

A.9 Rankings of states with variable costs (Section 6.1)

This appendix examines when Assumption 2 holds with no restrictions on the demand curve. Consider any pair of states, and call them good (G) and bad (B). FBO usage levels are determined by the respective conditions

$$p(N_G^o) = C_G(N_G^o) + C_G'(N_G^o) N_G^o, \tag{A.13}$$

$$p(N_B^o) = C_B(N_B^o) + C_B'(N_B^o) N_B^o. \tag{A.14}$$

FBO fees are $f_G^o = C_G'(N_G^o) N_G^o$ and $f_B^o = C_B'(N_B^o) N_B^o$.

Suppose the cost function in state $\omega \in \{G, B\}$ has the form

$$C_\omega(N) = c + d_\omega N^\varepsilon, \tag{A.15}$$

where $c \geq 0$ and $\varepsilon > 0$. Eqns. (A.13) and (A.14) become

$$p(N_G^o) = c + (1+\varepsilon) d_G (N_G^o)^\varepsilon,$$

$$p(N_B^o) = c + (1+\varepsilon) d_B (N_B^o)^\varepsilon.$$

Assumption 2(a) requires $d_B > d_G$. It then follows that $N_B^o < N_G^o$, $p(N_B^o) > p(N_G^o)$, and $f_B^o = \varepsilon d_B (N_B^o)^\varepsilon > f_G^o = \varepsilon d_G (N_G^o)^\varepsilon$. Assumption 2(b) is therefore satisfied as well. Note that the cost function does not have to be convex since parameter ε can be less than 1.

We now return to general cost functions and identify two necessary conditions for Assumption 2 to hold.

Condition 1 *The cost function must be steeper in less favorable states. With two states, G and B, the requisite condition is* $C_B'(N) \geq C_G'(N)$ *for all* $N > 0$.

To see that Condition 1 must hold, note that if the demand curve is nearly vertical then $N_B^o \cong N_G^o$ and the condition $f_B^o \geq f_G^o$ simplifies to $C_B'(N) \geq C_G'(N)$.

Condition 2 *Congestion-free costs must be the same in all states.*

To see this, consider a linear variant of eqn. (A.15):

$$C_G(N) = c_G + d_G N, \;\; C_B(N) = c_B + d_B N,$$

with $d_B > d_G$. Assumption 2(a) requires $c_B \geq c_G$. Suppose the demand curve is horizontal with a choke price of $\bar{p}$. First-best fees are

$$f_G^o = \frac{\bar{p} - c_G}{2}, \; f_B^o = \frac{\bar{p} - c_B}{2}.$$

Assumption 2(b) requires $f_B^o \geq f_G^o$, which implies $c_B \leq c_G$. Hence $c_B = c_G$.

A.10 Proof of Theorem 5 (Section 6.1)

We first prove Theorem 5 for cases in which the permit constraint binds in all states. We then show that the same reasoning applies if the constraint does not bind in some states. To facilitate the proof, we assume that the number of states is countable and finite. A similar proof applies if Ω is continuous.

List states from worst to best as per Assumption 2 so that for any states ω and $\omega + 1$, $N_{\omega+1}^o > N_\omega^o$ and $f_{\omega+1}^o < f_\omega^o$. Let k be the unique state such that $Y \in \left[N_k^o, N_{k+1}^o\right]$.

Lemma 1 *There exists a fee that is welfare-superior to the TPS for states k and $k+1$.*

Proof of lemma: Let f be the fee. Clearly, $f \in \left[f_{k+1}^o, f_k^o\right]$. A value of f within this interval that is welfare-superior to the TPS can be found by trial and error as follows. First set $f = f_k^o$. Since $f_{k+1}^o < f_k^o$, $N_{k+1}(f) < N_{k+1}^o$. If $N_{k+1}(f) \geq Y$ then $N_{k+1}(f) \in \left[Y, N_{k+1}^o\right)$. The fee supports the FBO in state k, and an outcome equal to or better than the TPS in state $k+1$. Suppose instead that $N_{k+1}(f) \leq Y$. Reduce f until $N_{k+1}(f) = Y$. In state $k+1$ the fee supports the same outcome as the TPS. Since $C_k(N) > C_{k+1}(N)$ for any $N > 0$, $N_k(f) < N_{k+1}(f)$. Moreover, since $f < f_k^o$, $N_k(f) > N_k^o$. Hence $N_k(f) \in (N_k^o, Y)$, and in state k the fee supports an outcome closer to the optimum than the TPS. Thus, the fee is welfare-superior to the TPS in both states k and $k+1$. ■

For the rest of the proof of Theorem 5 the fee is held fixed at the f identified in Lemma 1. Consider any state $j < k$. Given $C_j(N) > C_k(N)$, $N_j(f) < N_k(f) \leq Y$ so that $N_j(f) < Y$. Given $f \leq f_k^o < f_j^o$, $N_j(f) > N_j^o$. Hence $N_j(f) \in \left(N_j^o, Y\right)$, and usage in state j is closer to the FBO than with the TPS.

Now consider any state $j > k+1$. Given $C_j(N) > C_{k+1}(N)$, $N_j(f) > N_{k+1}(f) \geq Y$ so that $N_j(f) > Y$. With $f > f_j^o$, $N_j(f) < N_j^o$. Hence $N_j(f) \in \left(Y, N_j^o\right)$ and usage in state j is closer to the FBO than with the TPS. In conclusion, the fee is at least as efficient as the TPS in every state, and strictly superior in at least some states.

As a final step, suppose the TPS does not bind in all states. Since the unregulated usage level is strictly increasing with the state as per Assumption

2, it must not bind in states 1...j for some $j \geq 1$. The fee is set in the same way as when the permit binds in all states so that $f \in \left[f^o_{k+1}, f^o_k\right]$ for some state k. Usage in states k and higher is welfare-superior with the fee as before. Since $j \leq k$, $f \leq f^o_j$. The fee therefore reduces usage in states 1...j part way toward their respective FBO values $N^o_1 ... N^o_j$, and thus improves efficiency in these states. By contrast, the TPS does not improve efficiency in states 1...j at all. Thus, the fee is again at least as efficient as the TPS in every state, and strictly superior in at least some states. ■

A.11 Analytics of the adaptive tradable permit scheme (Section 7)

Let Ω_c denote the set of states in which the initial permit allocation prevails and the government neither buys nor sells permits. Let Ω_r denote the set of states in which the government buys permits, and Ω_s the set of states in which it sells permits. Usage in the three intervals is governed by the equations

$$\begin{aligned} p_\omega\left(N_\omega\right) &= C_\omega\left(N_\omega\right) + r, \ \ \omega \in \Omega_r, \\ N_\omega &= Y, \ \ \omega \in \Omega_c, \\ p_\omega\left(N_\omega\right) &= C_\omega\left(N_\omega\right) + s, \ \ \omega \in \Omega_s. \end{aligned}$$

With the linear model,

$$\begin{aligned} N_\omega &= \frac{1}{b+d}\left(A - r\right), \ \ \omega \in \Omega_r, \\ N_\omega &= \frac{1}{b+d}\left(A - q\right) = Y, \ \ \omega \in \Omega_c, \\ N_\omega &= \frac{1}{b+d}\left(A - s\right), \ \ \omega \in \Omega_s. \end{aligned} \tag{A.16}$$

Parameters Y, r, and s are chosen to maximize expected welfare in (11). The optimal permit allocation is

$$Y = \frac{\bar{A}_c}{b+2d}, \tag{A.17}$$

where $\bar{A}_c$ is the mean value of A for $\omega \in \Omega_c$. The first-order conditions for r and s are

$$r^* = \frac{E_{\omega\in\Omega_r}\left\{C'_\omega\left(N_\omega\right) N_\omega \frac{\partial N_\omega}{\partial r}\right\}}{E_{\omega\in\Omega_r}\left\{\frac{\partial N_\omega}{\partial r}\right\}}, \tag{A.18}$$

$$s^* = \frac{E_{\omega\in\Omega_s}\left\{C'_\omega\left(N_\omega\right) N_\omega \frac{\partial N_\omega}{\partial s}\right\}}{E_{\omega\in\Omega_s}\left\{\frac{\partial N_\omega}{\partial s}\right\}}. \tag{A.19}$$

Eqns. (A.18) and (A.19) have the same structure as eqn. (13) for the optimal fee.

Assume now that variable $A \equiv a - c$ is uniformly distributed on the interval $[A_0, A_1]$. Eqns. (A.18) and (A.19) simplify to

$$r^* = \frac{d}{b+2d}\bar{A}_r, \tag{A.20}$$

$$s^* = \frac{d}{b+2d}\bar{A}_s. \tag{A.21}$$

Let A_{rc} and A_{cs} denote the boundaries between sets Ω_r, Ω_c, and Ω_s so that $\Omega_r = [A_0, A_{rc})$, $\Omega_c = [A_{rc}, A_{cs}]$, and $\Omega_s = (A_{cs}, A_1]$. Then $\bar{A}_r = (A_0 + A_{rc})/2$, $\bar{A}_c = (A_{rc} + A_{cs})/2$, and $\bar{A}_s = (A_{cs} + A_1)/2$.

Setting $N|_{A=A_{rc}} = Y$ and using (A.16), (A.17), and (A.20) yields

$$(b+2d)\,A_{rc} = dA_0 + (b+d)\,A_{cs}. \tag{A.22}$$

Setting $N|_{A=A_{cs}} = Y$ and using (A.16), (A.17), and (A.21) yields

$$(b+2d)\,A_{cs} = dA_1 + (b+d)\,A_{rc}. \tag{A.23}$$

The solution to (A.22) and (A.23) is

$$A_{rc} = \frac{(b+2d)\,A_0 + (b+d)\,A_1}{2b+3d},$$

$$A_{cs} = \frac{(b+d)\,A_0 + (b+2d)\,A_1}{2b+3d}.$$

Substituting these formulas into eqns. (A.20) and (A.21) gives

$$r^* = \frac{d\,((3b+5d)\,A_0 + (b+d)\,A_1)}{2\,(b+2d)\,(2b+3d)},$$

$$s^* = \frac{d\,((b+d)\,A_0 + (3b+5d)\,A_1)}{2\,(b+2d)\,(2b+3d)}.$$

Note that in the limit $A_1 \to A_0$, $r^* = s^* = dA_0/(b+2d)$ which is eqn. (27) for the optimal fee, f^*.

The probabilities of the three states are

$$\wp(\Omega_r) = \frac{b+d}{2b+3d},\ \wp(\Omega_c) = \frac{d}{2b+3d},\ \wp(\Omega_s) = \frac{b+d}{2b+3d}.$$

The government is equally likely to buy as sell permits, and trading in either direction is more frequent than not trading.

Routine algebra yields eqn. (43) for the relative efficiency of the adaptive TPS. The relative efficiencies of the fee and the basic TPS are readily derived using eqns. (28) and (30), respectively.