Estimating Urban Road Congestion Costs

David Newbery
Georgina Santos (YE)

David Newbery
Department of Applied Economics
University of Cambridge, UK
Cambridge CB3 9DE
Tel: + 44 1223 33 52 47
Fax: + 44 1223 33 52 99
David.Newbery@econ.cam.ac.uk

Georgina Santos (YE)
Department of Applied Economics
University of Cambridge
Cambridge CB3 9DE, UK
Tel: + 44 1223 33 52 84
Fax: + 44 1223 33 52 99
Georgina.Santos@econ.cam.ac.uk

Estimating Urban Road Congestion Costs

David M Newbery and Georgina Santos* (YE)
Department of Applied Economics
Cambridge, UK
December 31 2001

Abstract
Economists wishing to analyse road congestion and road pricing have usually relied on link-based speed-flow relationships. These may provide a poor description of urban congestion, which mainly arises from delays at intersections. Using the simulation model SATURN, we investigate the second-best proportional traffic reduction and find that linear speed-flow relations describe network flows quite well in eight English towns, though the predicted congestion costs and charges overstate those apparently required in our second best model. We then confront the results with feasible optimal cordon charges, and find them reasonably correlated, but imperfect predictors.

Key words
Congestion tolls, traffic congestion, road pricing, efficient charges

JEL classification
H54, H11, R41, R48.

Introduction
Economists wishing to analyse road congestion and road pricing have usually relied on link speed-flow relationships estimated by traffic engineers, typically piece-wise linear, to compute marginal congestion costs (e.g. Newbery, 1988, 1990).

One obvious objection is that observations of traffic flows on links may be a very poor guide to traffic conditions in densely meshed urban networks, where most traffic interactions and delays take place at intersections, not on the links between. This in turn casts doubt on estimates based on these measured simple speed-flow relationships.

The aim of this paper is to draw on the results of urban network traffic simulation models to test the reliability and applicability of linear speed-flow relationships, and compare their implied congestion costs with more soundly based measures. In so doing we have to confront the obvious difficulty that it is infeasible to perfectly reflect marginal congestion costs on each part of the network and each time of the day in a set of time-varying, location-specific charges. Instead we need to address the question of how congestion pricing might be implemented in practice, and what relation the simplified or second-best congestion charges might bear to those we can estimate from our simple network relationships. We do this in the second part of the

* Support from the ESRC under Grant R000223117 *Road pricing and urban congestion costs*, and from the Department of the Environment, Transport and the Regions (DETR), under Contract N° PPAD 9/99/28, is gratefully acknowledged, as is support for Georgina Santos from the British Academy. Any views expressed in this paper are not necessarily those of the ESRC, DETR, British Academy, or University of Cambridge. The authors are grateful to Prof. Dirck Van Vliet of the Institute for Transport Studies at University of Leeds for his kind patience in clearing their doubts about SATURN. We are indebted to James Lindsay, from WS Atkins, who provided us with data on vehicle counts in Cambridge and to Markus Kuhn for support and advice on batch file programs.

paper, which estimates the costs and benefits of feasible cordon tolls for eight English towns. Cordon tolls have been implemented successfully in Singapore and in a number of cities in Norway and a scheme of this sort is expected to be in place in London by 2003.

While these results are interesting in their own right, we are also interested in assessing how well simple models do in predicting the benefits of this form of feasible road charging. If they are, it should be possible to screen towns to identify promising subjects for more detailed investigation. If not, then time-consuming simulation approaches will be required. Unfortunately, not all suitable towns have such models, which are very costly to calibrate.

The standard approach

Figure 1 illustrates the traditional analysis. Higher traffic flows lead to lower average speeds and higher travel times and costs per km. Additional traffic imposes an external cost on all other road users. Under congested conditions, particularly in urban areas, and in the absence of efficient road pricing, traffic will be undercharged and hence excessive.

In Figure 1, the average social cost (ASC) excludes road taxes, while the average private cost (APC) includes road taxes, CB. If the inverse demand can be represented as shown, then the equilibrium will occur at point C, where the marginal willingness to pay is equal to the APC. The efficient equilibrium is at point D, where the marginal social cost (MSC) is equal to the marginal willingness to pay. This would be supported by a congestion charge DE (which, as shown, would replace the poorly targeted road tax levied on fuel, CB). The inefficiency of incorrect pricing is then measured by the area DCM.

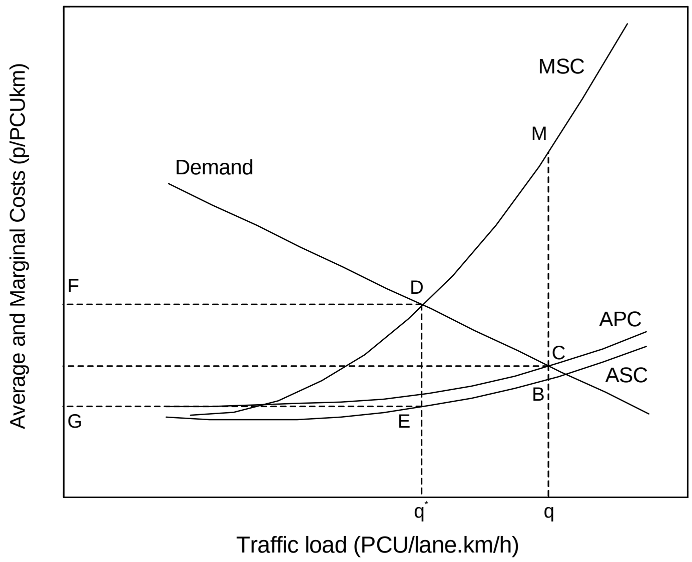

Figure 1: Average and Marginal costs, deadweight loss, efficient charge

Algebraically, if *c(q)* is the ASC when the traffic flow is *q*, and $C = cq$ is total social cost per km of the traffic flow, then the MSC is dC/dq, while the APC is $c + \tau$, where τ is the road tax per km. If *p(q)* is the demand price at traffic flow *q* (and associated private cost), then the deadweight loss (DWL), area DMC, is the area between the marginal social cost ADM and the inverse demand curve DC from their point of intersection, q^*, at D in Figure 1 (the efficient level of demand) and the actual level of demand, *q*, at C. This is given by

$$L = \int_{q^*}^{q} \left(\frac{dC}{dq} - p(q) \right) dq \approx \Delta C(q) - \Delta q (\bar{c} + \tau) , \tag{1}$$

where $\Delta q = q - q^*$, $\Delta C = C(q) - C(q^*)$, $\bar{c}$ is the average of the (pre-tax) unit costs at the efficient and actual levels of demand, and the approximation derives from assuming that the demand for travel is approximately linear in private cost. Alternatively, the DWL triangle is approximately $1/2.\Delta q.e$ where e is the externality or marginal congestion cost (MCC), at the market equilibrium, shown as MC in Figure 1.

The formula for c (and hence for C) is

$$c = a + b/v , \tag{2}$$

where a is the cost per PCU.km (pence/PCU.km),[1] and b is the value of time (pence/PCU.h). The simplest speed-flow relationship is (piece-wise) linear, which over the relevant range (around the observed equilibrium) is

$$v = v_0 - \beta q . \tag{3}$$

The MCC, e, per PCU.km is the marginal social cost, dC/dq, less the private cost, c, and by differentiating (2) and (3) is

$$e \equiv \frac{d}{dq}(cq) - c = q\frac{dc}{dq} = \frac{b\beta q}{v^2} = \frac{b(v_0 - v)}{v^2} . \tag{4}$$

The US Bureau of Public Roads uses a different formula for the time to travel on link i, t_i:

$$t_i = t_{i0}\left(1 + \gamma [q_i / k_i]^4 \right), \tag{5}$$

where t_{i0} is the time with no traffic, k_i is the capacity of the link, and $\gamma = 0.15$ (Anderson and Mohring, 1996). Using the same methodology as above, the per vehicle externality on link i is can be expressed as

$$e_i = b q_i \frac{dt_i}{dq_i} = 4b\, q_i \gamma\, t_{i0} \frac{q_i^3}{k_i^4} = 4b(t_i - t_{i0}) . \tag{6}$$

Expressing this as an externality per PCU.km, as in (4), gives the MCC

$$e = \frac{4b(v_0 - v)}{v\, v_0} = \frac{b(v_0 - v)}{v^2} . \left(\frac{4v}{v_0} \right). \tag{7}$$

Comparing this with (4), the externality will be higher unless the speed falls below one-quarter the intercept speed, v_0. The cautionary lesson to draw is that the choice of the speed-flow functional form may pre-determine the congestion externality costs. Functional forms that

[1] Traffic flow is measured by its congestive effect in PCU or Passenger Car Units, where 1 car = 1 PCU, but a bus in urban conditions may be 3 PCU.

appear to fit equally well over the range of observations may give rise to significantly different congestion externality costs away from the range of observation. The problem is that the size of the externality cost depends on the elasticity of the speed flow function,[2] whose fitting is more likely to concentrate on estimating either the average slope of the function (where this is piecewise linear) or the parameters of some class of functions. For our immediate purposes, though, the value of equation (7) is that it roughly calibrates v/v_0 as 1/4.

Implications of the linear speed-flow relationship

If the speed-flow relationship is linear over the relevant range (from the efficient to the market equilibrium level of traffic, or from D to C in Figure 1), as in (3), then it is relatively simple to solve for the efficient level of traffic. Let unsubscripted variables refer to the efficient levels of those variables, and subscript *m* refers to the market equilibrium, thus c_m. The efficient price, *p*, is equal to the average price cost (APC) *plus* the MCC, *e*: $p = c + e$.[3] If demand is linear, and has elasticity ε at the original market equilibrium, the price at the efficient equilibrium is given by

$$p = c_m\left(1 + \frac{1}{\varepsilon}\left(\frac{q_m - q}{q_m}\right)\right) = c + e. \tag{8}$$

This can be written as an equation in *v* by replacing q/q_m by $(v_0 - v)/(v_0 - v_m)$ from (3), from which q/q_m can be recovered:

$$\left(\frac{a}{b} + \frac{1}{v_m}\right)\left(1 + \frac{1}{\varepsilon}\frac{v - v_m}{v_0 - v_m}\right) = \frac{a}{b} + \frac{v_0^2}{v^2}. \tag{9}$$

This is now a cubic equation in *v*, which can be solved numerically. It can also be written in terms of parameters whose values can be readily approximated. Thus if $av_0/b = \theta$, the ratio of per km costs to the time cost at the 'intercept' speed, v_0, $v_m/v_0 = \varphi$ (which, from (7) will be about 1/4), and $v/v_0 = x > \varphi$, then (9) is a cubic equation in *x*:

$$\left(\frac{1 + \theta\phi}{\phi}\right)\left(1 + \frac{1}{\varepsilon}\frac{x - \phi}{1 - \phi}\right) = \theta + \frac{1}{x^2}. \tag{10}$$

Similarly, the ratio of the required charge to the APC, *e/c*, is a function of *x*, while the price of travel with the charge may also usefully be compared with the original APC:

$$\frac{e}{c} = \frac{1 - x}{x + \theta x^2}; \quad \frac{p}{c_m} = \frac{\theta + 1/x^2}{\theta + 1/\phi}. \tag{11}$$

For example, if $\varepsilon = 2/3$, $\theta = 1/2$, $\varphi = 1/4$, then $x = 2/5$, $v/v_m = x/\varphi = 1*3/5$, $e/c = 5/4$, $p/c_m = 2/3$, and $q/q_m = 4/5$. The charge required is large (it increases the cost of travel by 50%), but the increase in speed is large (60%), implying large time savings (a reduction in trip time of 37.5%), for a required reduction in traffic of 20% and the charge needed. If the elasticity is lower, say $\varepsilon = 0.1$, then $x = 1/3$, $v/v_m = 1*1/3$, $e/c = 1*5/7$, $p/c_m = 2*1/9$, and $q/q_m = 8/9$. The average travel time

---

[2] From (4), the ratio of the congestion externality to the time cost, *e/(bt)*, is the elasticity of the speed-flow relationship, $-d \log v / d \log q$.

[3] For simplicity, the distinction between ASC and APC, and the existing fuel tax per km, will be ignored.

still drops by one quarter for a more modest efficient reduction in traffic of 11%, though the charge needed to achieve this improvement is considerably higher. With a higher elasticity of $\varepsilon = 0.7$, $x = 0.426$, $v/v_m = 170\%$, $e/c = 1.11$, $p/c_m = 1.34$, and the reduction in trip time is 41% for a traffic reduction of 23% and a more modest charge. These are surprisingly large changes from the market equilibrium.

If the demand schedule has constant elasticity, ε, then the equation for x is more complex, but still numerically soluble:

$$p = c_m \left(\frac{q}{q_m} \right)^{-1/\varepsilon} = c + e\,. \tag{12}$$

Substituting as before gives the following equation:

$$\left(\frac{a}{b} + \frac{1}{v_m} \right) \left(\frac{v - v_m}{v_0 - v_m} \right)^{-1/\varepsilon} = \frac{a}{b} + \frac{v_0^2}{v^2}, \text{ or } \left(\frac{1+\theta\phi}{\phi} \right) \left(\frac{1-x}{1-\phi} \right)^{-1/\varepsilon} = \theta + \frac{1}{x^2}\,. \tag{13}$$

Thus if $\varepsilon = 2/5$, $\theta = 1/2$, $\varphi = 1/4$, $x = 0.382$ (slightly lower than in the linear case), and $q/q_m = 0.82$, slightly higher, so congestion is reduced rather less. The charge required is higher ($e/c = 1.35$ instead of 1.25, $p/c_m = 1.63$ instead of 1.5), and trip time falls by 35% instead of 37.5%. At an elasticity of 0.2, $e/c = 1.62$, $p/c_m = 1.97$; traffic falls by 12.7% and trip times fall by 27.6%, while at $\varepsilon = 0.7$, $e/c = 1.18$, $p/c_m = 1.42$, traffic falls by 21.7% and trip times by 39.4%, quite close to the linear approximation. The reason for these systematic differences is that the linear demand schedule has a higher elasticity at the efficient equilibrium than the constant elasticity demand schedule against which it has been compared.

It is also easy to estimate the deadweight loss per km of the market equilibrium from the triangular approximation $L/q_m = 1/2.\Delta q.e_m/q_m$, where $\Delta q = q_m - q$. This can be compared with the APC, c_m, or the revenue per km from efficient road pricing, e:

$$\frac{L}{q_m c_m} = \frac{1}{2} \frac{x-\phi}{\phi^2\theta + \phi}, \quad \frac{L}{q\,e} = \frac{1}{2} \frac{(1-\phi)(x-\phi)\,x^2}{\phi^2 (1-x)^2}\,. \tag{14}$$

Thus using the above parameters, for linear demand with an elasticity of 0.4, $L/c_m.q_m = 4/15 = 0.27$, $L/e.q = 2/5$.

Modelling urban congestion

Simple link-based models relating speed to flow fail to capture the complex network interactions that occur in even moderately sized congested towns. Fortunately, sophisticated traffic assignment models have been developed to simulate equilibrium traffic flows over a network. They simulate the results of demands for trips specified by a matrix giving the number of trips between all origin and destination (O-D) pairs. We have chosen to use SATURN (Simulation and Assignment of Traffic to Urban Road Networks), a software package developed at the Institute for Transport Studies at Leeds University, that allows us to compute the costs of vehicle trips for varying levels of traffic (Van Vliet and Hall, 1997). SATURN finds the Wardrop equilibrium, defined as the assignment of traffic in which no trip-maker can reduce his or her total trip cost (the value of the time taken and vehicle operating costs) by choosing a different route to that assigned. The model can compute the total cost of the market equilibrium traffic flow, once the model has been calibrated to the town and traffic in question. It can also compute the costs of variations around this equilibrium, though the results become less reliable the further from the observed flows. This is not only because of the inherent difficulty of accurately modelling very different demands, but also because the traffic management system (traffic lights, priorities, parking restrictions, etc) would be adjusted to the observed pattern of flows in ways that are hard to predict.

Interpreting the deadweight loss diagram

Although Figure 1 is useful in illustrating the concepts, and is standard for representing congestion in the literature, it conceals a number of hidden assumptions and problems. The first problem is that demand is for trips between origin-destination (O-D) pairs, while the costs are specific to particular links along that trip. Unless all links are equally congested, or all traffic travels between the same origin and destination, there is a mismatch between the entity demanded and the costs of the supply.

There are a number of ways of addressing this mismatch between demand and supply, all of which encounter either conceptual or computational difficulties. The theoretically correct but computationally demanding approach would be to compute the marginal congestion costs on each link and junction. (Most of the congestion actually occurs at junctions, and depends on the detailed geometry of the junction as well as flows on other arms, as well as that on the arm used by the vehicle in question.) From this one could determine the appropriate congestion charge for each link and junction, apply this set of charges and re-compute the total cost of various trips. Different routes will now have different costs (now that they include a congestion charge for each link and junction) and in response demand for different trips will fall by varying amounts, and probably the pattern of routes taken will also change.

Possibly a model could be developed to iterate towards a new equilibrium. Then for each O-D pair, the social costs and consumer benefits could be computed. Unfortunately, SATURN is not configured to compute this new equilibrium, and we need to find a practical approximation to estimate the DWL associated with the original inefficient market equilibrium. It is important to understand their limitations if the resulting numerical results are to be correctly interpreted.

Another obvious limitation of Figure 1 is that it only applies when traffic can be measured by a scalar variable (in the figure, traffic flow in PCU/lane.km/h). If the diagram is to be interpreted for urban road networks, we need some method of converting the complex pattern of flows on links into a scalar measure of overall traffic. Fortunately, SATURN provides a simple method of scaling flows by changing the number of trips between each origin and destination (O-D) pair in proportion. This is the only natural scaling method available, and although it has obvious limitations, which will be discussed below, it allows natural interpretations of the concepts of the average marginal congestion cost and the estimated value of the DWL. We can then return to the more fundamental question of how this measure relates to the true inefficiency of congestion, and, more important, to the measurement of potential benefits from feasible systems of road charging.

We start by numbering the set of O-D trips by i, where $i = 1,2,3,...m.n$, if there are m origins and n destinations. The next step is to work in terms of the units demanded, namely trips, rather than PCU.km on specific parts of the network. To that end, we define the following notation, using capitals for trip-related variables:

- the social cost for the trip i is S_i pence/PCU,
- the private cost of trip i is P_i pence/PCU (the perceived effective total cost of time and distance),
- the length in km of trip i is d_i km,
- the number of trips i is Q_i PCU, $Q_i = Q_i(P_i)$,

- the consumer *utility* (measured in cash terms) of the total number of trips i is $U^i(Q_i)$, where $dU/dQ = f_i(Q_i) = P_i$ in equilibrium.

If the level of traffic relative to the equilibrium (subscript 0) is measured by the scalar θ, then the actual number of OD trips i is $Q_i = \theta Q_{i0}$ for all i. The consumer utility can be also expressed as an integral of the inverse demand schedule:

$$U_i(Q_i) = U_i(\theta Q_{i0}) = \int_{\underline{Q}}^{\theta Q_{i0}} f_i(Q)dQ\,, \tag{15}$$

where $\underline{Q}$ is some arbitrary but fixed minimum level, possibly zero.

Social welfare at traffic level θ will be $W(\theta) = \Sigma U^i(\theta Q_{i0}) - C(\theta)$, where $C(\theta)$ is the total social cost of trips, $\Sigma\theta Q_{i0}.S(\theta Q_{i0})$. The uniform reduction that maximises social welfare (the second best optimum)[4] can be found by differentiating $W(\theta)$ (using the integral form of consumer utility):

$$\frac{dW}{d\theta} = 0 \Leftrightarrow \frac{dC}{d\theta} = \sum_i Q_{i0}\, f_i(Q_i) \tag{16}$$

To relate this to the various curves in Figure 1 we need to express prices and costs in pence/PCU.km. The total PCU km travelled (PCUKT) on trip i is d_iQ_i, and total PCUKT is $D = \Sigma d_iQ_i$. The average private cost, a, and the marginal social cost, MSC, m, are then:

$$a = \frac{\sum_i p_i Q_i}{D}; \quad m = \frac{d}{dD}\left(\sum_i c_i Q_i\right). \tag{17}$$

To a close approximation,[5] $D = \theta D_0$, where $D_0 = \Sigma Q_{i0}d_{i0}$, so at the optimum

$$m = \frac{dC}{dD} = \frac{1}{D_0}\frac{dC}{d\theta} = \frac{\sum_i Q_{i0}\, f_i(Q_i)}{\sum_i Q_{i0}\, d_{i0}}. \tag{18}$$

It remains to define the average price such that at the optimum, the MSC is equal to the price. Fortunately, the condition will be satisfied with the natural interpretation that the average price is the trip-weighted price per PCU.km:

$$P = F(\theta) = \frac{\sum_i Q_i\, f_i(Q_i)}{D} = \frac{\sum_i \theta Q_{i0}\, f_i(Q_i)}{\theta D_0} = m\,, \tag{19}$$

as required for (second-best) optimality. It also follows that the deadweight loss is correctly measured by the area between the demand schedule, the MSC schedule, and the vertical line through the market equilibrium at $\theta = 1$.

---

[4] Second best because all trips are constrained to be equiproportionately reduced, rather than all trips being varied to maximise social welfare.

[5] As traffic decreases, the average distance travelled per trip may decrease slightly through re-routing. The effect is quantitatively small.

The final question is whether the required tax in Figure 1, $P(\theta)$-$a(\theta)$, is the average of the taxes required to reduce each trip to a fraction θ of its original level. The tax required on trip i is $t_i = f_i(Q_i) - P_i(Q_i)$, to the average tax per PCU.km, t, is given by:

$$t = \frac{\sum_i t_i Q_i}{D} = \frac{\sum_i Q_i f_i(Q_i)}{D} - \frac{\sum_i Q_i P_i(Q_i)}{D} = m - a \,. \tag{20}$$

Consequently, the measured average corrective tax derived in Figure 1 is equal to the road charges that would have to be levied on each trip to reduce demand by the desired amount. Note that these road charges would have to be levied solely as a function of origin and destination to avoid influencing the choice of route (other than in response to the changed level of traffic). Such road charges are doubly inefficient relative to the first best, in that they do not discourage traffic from the most congested parts of town, nor do they penalise trips with higher external costs more heavily than those with low external costs. They do, however, allow one to place meaning on the various costs, externalities and corrective charges in Figure. 1. We can summarise these various results in the following proposition:

Proposition: *The second-best optimum in which all trips are reduced proportionately can be supported by trip-based tolls whose trip-weighted average value would raise the average private cost to a price on the aggregate trip demand equal to that optimum level. The measured DWL is then an accurate measure of the gains from this uniform reduction*.

Calculating second-best congestion costs for eight towns

We have obtained and mounted calibrated SATURN network and traffic (O-D) files for eight towns: Cambridge, Northampton, Kingston upon Hull, Lincoln, Hereford, Bedford, Norwich and York. The model was run for the morning peak from 8 to 9 am. The vehicle operating cost, a in (2), and the value of time, b in (2), were taken as 12 pence per PCUkm and 1400 pence per hour respectively (1998 prices). These values were computed as weighted averages taking into account vehicle and fuel type, vehicle occupations, trip purpose and average wages and value of leisure time, according to guidelines of the Highways Economics Note N°2 (Highways Agency *et al*, 1996).

By considering a range of values of the scaling parameter, θ, from 0.05 to 1.50, the values for total cost and hence APC, MSC and MCC could be determined. Given a specification of demand, the second best optimum level of trips, MCC and hence average toll per km could be estimated, together with the DWL. Various demand specifications were tried, and for the same average elasticity gave very similar results, so only the constant elasticity demand specification will be reported here. Two values for the elasticity, spanning the plausible range of values for short-run demand responses in urban congested areas, were used: 0.2 and 0.7 (expressed as positive numbers).

Most towns include all trips, some of which originate some distance from the congested network.[6] As a result, a typical trip will be a combination of high-speed travel on relatively uncongested trunk roads, and low speed travel on congested urban streets. In order to concentrate attention on the congested urban network, the models have been reconfigured such that trips notionally start on the boundary of a defined area, either just inside the outer ring roads, or in the central area. For most towns we therefore have three possible sets of data to analyse: for the whole urban area, including surrounding motorways and trunk roads, for the urban area proper, excluding any surrounding motorways and trunk roads, and for the central

[6] Cambridge includes trips from very distant villages, and has been truncated to mimic the effect of trips starting close to the outer ring road, which lies in the green belt.

zone or city centre. The simulations also allowed us to estimate the average network speed (total distance travelled divided by total time) for each value of θ. Finally, the fitted linear speed-flow relationship was estimated by ordinary regression over the range from just below the properly computed second best optimum to the market equilibrium traffic level. The linear relation fits reasonably well over this range, though the range of linearity is restricted to not more than this range, as Figure 2 demonstrates.

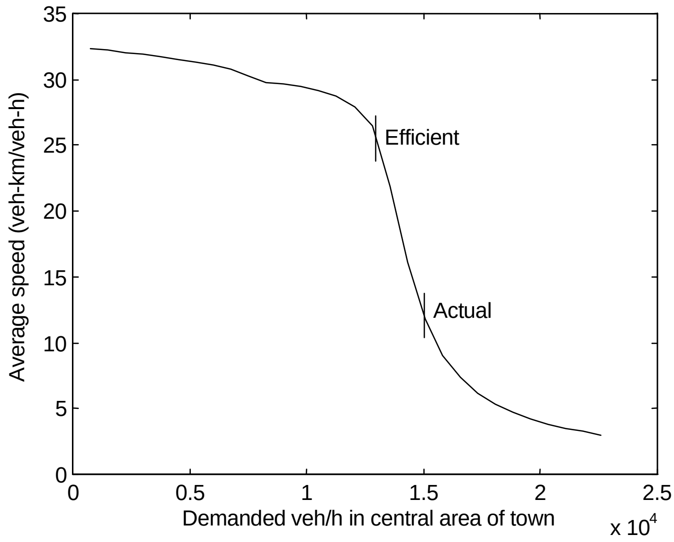

Figure 3: Speed Vs demanded flow in Central Northampton during the morning peak

The critical parameters for the eight towns are given in Table 1. Note that the intercept value v_0 depends on the range of traffic values over which the relation is fitted, and for higher elasticities, the efficient equilibrium will be lower, and further from the observed equilibrium. Typically the wider the range of traffic values, the lower is the value of v_0. This itself is of some concern, as most of the computed values of interest are sensitive to this parameter. Thus the marginal congestion cost, MCC, computed from (4), typically exceeds that computed using SATURN (compare the lines "MCC pred" and "MCC SAT"). This in turn has a considerable effect on the DWL, which is estimated from $1/2.\Delta q.e_m$, so the error in the estimated DWL is directly proportional to the error in MCC, even if the efficient traffic volume is correctly predicted. Fortunately, the latter appears to be the case, though the average charge required to sustain that is also overestimated.

Table 1 computes the efficient level of traffic and compares two methods of estimating congestion costs for the same set of trips.

Table 1: Comparisons of congestion from linear speed-flow relationships in whole towns. Elasticity: 0.2

Town	Northampton	Kingston upon Hull	Cambridge	Norwich	Lincoln	York	Bedford	Hereford	Average
vo kph	110.8	88.7	112.7	50.9	83.0	82.1	85.7	79.6	86.7
vm kph	16.4	21.3	36.6	35.9	30.7	33.7	57.0	30.8	32.8
MCC pred	495	209	80	16	78	60	12	72	128
MCC SAT	315	166	71	14	67	44	11	57	93
eff. traff pred	0.87	0.87	0.89	0.95	0.89	0.90	0.95	0.90	0.90
eff. traff SAT	0.86	0.88	0.89	0.92	0.91	0.92	0.92	0.91	0.90
eff spd v kph pred	29	30	45	37	36	38	58	36	39
eff spdv kph SAT	29	30	48	36	36	38	58	35	39
charge pred	223	123	55	15	57	46	11	54	73
charge SAT	141	84	42	11	42	33	9	38	50
DWL pred	33	13	4.4	0.4	4.1	2.9	0.3	3.7	7.7
DWL SAT	23	10	3.9	0.3	3.4	2.0	0.2	2.7	5.7

Note: vo kph: intercept (kilometres per hour), vm kph: speed computed with SATURN or market speed (kilometres per hour), MCC pred: Marginal Congestion cost (predicted), MCC SAT: Marginal Congestion Cost computed with SATURN results, eff. traff pred: efficient traffic predicted, eff. traff SAT: efficient traffic computed with SATURN results, eff spd v kph pred: efficient speed (predicted), eff spd v kph SAT: efficient speed (from SATURN results), charge pred: charge predicted, charge SAT: charge SATURN, DWL pred: deadweight loss predicted, DWL SAT: deadweight loss SATURN

Assessment of area linear speed-flow relationships

The linear relationships give reasonable predictions for the efficient level of traffic and the efficient speeds, but are less accurate in predicting the MCC (and hence the DWL and efficient charge required). This should not be surprising, as the latter are sensitive to the local curvature of the relationship, while the former are less sensitive. What is striking is the wide variation across towns, which is far larger than the errors in predicting values for any one town. That suggests that observations on average traffic speed may provide valuable information about congestion costs and even the desired reductions needed.

Feasible road pricing schemes

The calculations reported above are in a sense the wrong answer to the wrong question. Whether or not urban traffic congestion is a serious problem depends partly on the size of the resulting inefficiency, which we have attempted to measure through an approximate though incorrect measure of the DWL. The measure is approximate in that it only considers uniform reductions of all trips. On the other hand, the important question is how large are the benefits that can be achieved by feasible traffic management or road pricing schemes. These charges are likely to be considerably cruder than charging each trip its marginal congestion cost, and thus would not achieve the theoretically possible (first best) benefits. Thus feasible schemes will have lower benefits than the first-best potential benefits, which exceed those computed from our second best uniform reduction.

The next question to ask is whether the simple computations reported above give any guidance on the likely benefits from feasible road pricing schemes, and whether the likely benefits are indeed greater than the costs. To that end we estimate the costs and benefits of cordon tolls for our eight towns.

In a cordon toll scheme a trip maker is charged a fixed amount to enter and/or leave the charged area at all or only some times of the day. The physical location of the roadside sensors determines the boundary of the charged area and defines the cordon. We based the decision of where to put the cordon on two main considerations: it should contain the most congested area, and not allow too many alternative routes. Only inbound cordons were considered.

The program used to estimate the tolls was SATURN together with a batch file procedure that simulates cordon tolls, SATTAX. SATTAX was developed by David Milne, also at the Institute for Transport Studies at Leeds University, and can be added to SATURN in order to simulate road charging (Milne and Van Vliet, 1993). SATTAX simulates a toll as a time penalty for crossing the cordon. The time penalty required will depend on the value of time assumed (23.4 pence per PCU.minute at 1998 prices). Thus a toll of £2.5 per crossing would be modelled as a delay of 641 seconds.

SATTAX allows for two kinds of responses: route choice and transfer off the road. Transfer off the road includes all trips that for one reason or other are dropped from the original trip matrix for the time period under study. The reasons for these trips to be excluded include change of departure time, change of mode, car pooling, and cancellation of the trip.

Results of cordon tolls

The criterion used to assess the benefits from a cordon toll was the increase in social surplus. We define social surplus as the trip makers' surplus, defined as the sum of individual utilities *less* total social cost. In the case of a unique origin-destination pair, the utility of driving is the integral under the inverse demand function between some reference level and the actual level of traffic. The difference between *ij* drivers' utility before and after the introduction of the toll was computed. That was done for each origin-destination pair and then all the changes in utilities were added up to get the overall change in utility. The change in total costs was obtained directly from the new cost matrix produced by SATTAX, adjusted to exclude VAT and fuel duties.

SATTAX was used to find the optimal toll, defined as the toll for which the social surplus reaches a maximum, as shown on Figure 3. Figure 3 also shows the effects of different elasticities - the higher the elasticity the higher the gain at any toll level. The main results are presented in Table 2. The annual gross revenues were computed as the number of vehicles that would cross the cordon multiplied by the toll that they would pay and by the number of working days (assumed to be 250) per year.

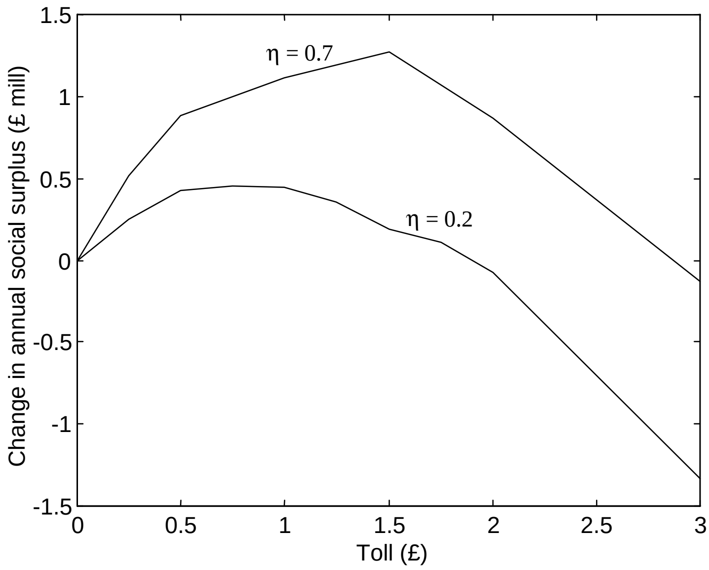

Figure 3: Change in annual social surplus in £ million at 1998 prices for different tolls and elasticities in Cambridge

Table 2. Optimal cordon tolls

Town	Optimal toll (£ to cross the cordon)	Benefit (£ mill./year)	Gross Revenues (£ mill./yr)	Ratio Revenue: Benefit
Northampton	3.00	2.37	8.25	3.5
Hull	2.50	3.24	7.69	2.4
Cambridge	0.75	0.45	1.75	3.9
Lincoln	0.25	0.44	0.52	1.2
Norwich	0.50	0.90	1.24	1.4
York	0.75	0.72	1.21	1.7
Bedford	0.50	0.52	1.21	2.3
Hereford	3.50	0.53	4.26	8.0

Source: Santos, Newbery and Rojey (2001)
Note: elasticity assumed: 02.

Cost-benefit analysis

Costs

All the costs were taken from Cheese and Klein (1999). The intra-vehicle unit's (IVU) implementation costs were assumed to be £15. Infrastructure costs were assumed to be

£45,300 per point[7]. One fourth of the cordon points were assumed to be dual lane, and would therefore require gantries. The cost of one gantry was assumed to be £97,000.

Operating costs, which include labour costs, costs of maintenance and costs of operating the infrastructure, were estimated using data from the Norwegian Public Roads Administration. These are in the order of 7 pence per transaction at most. The IVUs were assumed to have a life of six years, the electronic devices in the infrastructure five years and a value of £18,765 per cordon point (Cheese and Klein, 1999). The rest of the infrastructure was assumed to have the same life as the equipment and the scheme was assumed to last 30 years. Details on cost calculation are described in Santos and Newbery (2001).

Benefits

The benefits are simply the increase in social surplus. Details of the calculations are described in Santos and Newbery (2001). The increase in social surplus for a whole day was assumed to be three times the increase in social surplus from 8 to 9 AM. This is a conservative but reasonable assumption. The inefficiency is almost as high during the evening peak as during the morning peak (Newbery and Santos, 1999). Benefits increase with the elasticity, so at the assumed value of 0.2 they are likely underestimated. If a scheme is worthwhile with this value, then it will certainly be at higher and possibly more reasonable values.

Comparison of costs and benefits

Cordon tolling would be worthwhile only if the net present value (NPV) of benefits less costs were positive. Revenues are transfers, not benefits, and should not be part of the cost benefit analysis, though they are clearly of central interest to the charging authority and are the mechanism by which the costs are covered. There are additional benefits linked to the reduction in emissions, discussed below, but not included in these estimates of NPV. In addition, there may be a benefit from fewer accidents. We have ignored this benefit, as it would require further study to reach firm conclusions.

If there are distortions elsewhere in the urban economy, there would be a case for extending the analysis to value the impacts of transport changes on these distorted sectors, but this would be a major undertaking in its own right and not one we have considered. Table 3 summarises the costs and benefits for cordon tolls in the eight towns.

Table 3: Net Present Value of a cordon toll in different towns (£ million at 1998 prices)

Town		Total cost	Benefit	Net Present Value	Benefit/Cost
Cambridge	*Inner cordon*	16.1	17.6	1.5	1.1
	Two cordons (£1.25 each)	22.3	63.6	41.4	2.9
	Two cordons (£0.5 & £4.75)	22.3	101.8	79.5	4.6
	Outer cordon	15.9	89.6	73.7	5.6
Northampton		20.6	90.0	69.4	4.4
Kingston upon Hull		21.3	123.5	102.2	5.8
Hereford		9.7	19.5	9.8	2.0
Lincoln		14.5	17.0	2.4	1.2
Bedford		15.6	19.9	4.3	1.3

Source: Own calculations.

[7] MVA (1995) estimates infrastructure costs at £110,000 per point. Cheese and Klein's (1999) estimate was chosen instead because it is more recent and prices for this type of equipment are likely to decrease with time and technological progress.

The 1998 Treasury test discount rate of 6% was used.[8] A comparison of costs and benefits indicates that road pricing would be very beneficial in Kingston upon Hull and Northampton, and to a lesser extent in Hereford and York, on our conservative estimates. These schemes would become more beneficial if costs proved to be lower or if the elasticity of the demand proved to be higher than 0.2.

Two cordons in Cambridge, one inner and one outer, each charging £1.50 per crossing, yield a benefit-cost ratio almost three times as high that of an inner cordon scheme, thus making it attractive. A single outer cordon charging £5 increases the benefit-cost ratio to more than five times that of a single inner cordon. An outer cordon charging £4.75 combined with an inner cordon charging £0.50 yields higher benefits than an outer cordon implemented on its own. The costs are also higher because there are two cordons instead of one, but the net benefit is still increased, making this the preferred option.[9]

This shows that where a single inner cordon scheme might not be worthwhile, relocating the cordon may dramatically alter the benefits. In some cases a double cordon with each charge optimally set, might do even better (though if the tolls are not carefully set, much of the benefit may be lost). Varying the location and possibly the number of cordons is therefore likely to be worth investigating before deciding on the desirability of a road pricing scheme in any one town, for even where the viability of a single cordon is not in doubt, the benefits of one at a different location, possibly combined with an additional cordon, may greatly improve the outcome.

How well does the simple estimate of DWL do in predicting the benefits from cordon tolls? Table 3 shows the annual benefits of cordon tolls, and the predicted annual DWL using the linear speed-flow relations for the whole town. That would in principle be an indicator of the benefit that would be obtained from the introduction of an average charge. If these DWL estimates are compared with the benefits that would derive from a cordon toll scheme the following conclusions can be drawn:

a) Northampton and Kingston upon Hull, the towns with highest benefit/cost ratios (see Table 3), are, as expected, the towns with highest predicted DWLs (see Table 1).
b) The predicted DWL is not a very good estimate of the potential benefits of a cordon toll scheme. In most cases it underestimates them, with results that can be 80% or even 90% below the benefits that would accrue from the introduction of a cordon toll.
c) The estimate of DWL however can serve as a starting point when deciding whether to conduct a cost-benefit analysis to evaluate the possibility of implementing a road-pricing scheme in a town. Towns with high DWLs may be good candidates for cordon tolls, whereas towns with low DWLs would not probably benefit from a cordon toll (ie, the costs of introducing and operating such a system would be higher than the increase in social surplus that would derive from it).

Conclusions

The paper sets out to see how far simple speed-flow relations can be calibrated for urban networks and then used to predict congestion tolls, desirable traffic reductions and the resulting benefits from efficient road pricing. Perhaps surprisingly, given the radically different nature of congestion on links, for which the relations were first estimated, and in urban networks, the linear relationships do remarkably well. They describe the relationship between average speed and uniform scaling of all traffic, and as such give a reasonable estimate of the second-best extent of efficient traffic reduction - second best as it is the beset *uniform* reduction. The

---

[8] The Treasury was, in early 2001, reconsidering the test discount rate and may reduce it somewhat. If so, the benefit-cost ratio would be increased.

[9] In subsequent work, we found that relocating the inner cordon greatly increased the benefit, emphasising the point that the exact cordon location as well as the charge level can be critical for the success of the scheme.

estimated DWL (the benefits from that reduction) appear to overestimate the simulated DWL, and the required average congestion similarly overstates that needed.

This second best road pricing is, however, neither optimal nor feasible. It would be better to reduce traffic in proportion to congestion, not uniformly, and there is no feasible way of achieving either. Instead we considered another second best but feasible cordon toll scheme. For the eight towns costs and benefits were both computed, showing that for two towns, Northampton and Kingston upon Hull, the benefits considerably exceeded the costs, while in others the margin of benefit was uncomfortably small - Lincoln, Bedford, Norwich and Cambridge (with a poorly located single cordon, though the latter was dramatically improved by relocating the cordon and/or including a second cordon).

The DWL predicted with the linear speed-flow relations is not a very good indicator of the benefits that would be derived from a cordon toll scheme. As a rule of thumb however, towns with high DWLs may significantly benefit from a cordon toll scheme, with benefit/cost ratios well over 1.

References

Anderson, D. and H. Mohring (1996), *Congestion Costs and Congestion Pricing for the Twin Cities*, Report No MN/RC-96/32, Minnesota Department of Transportation, St Paul.

Cheese, J. and G. Klein (1999), *Charging Ahead: Making Road User Charging Work in the UK*, A Trafficflow Project Report, The Smith Group Ltd, Guildford, April.

Highways Agency, Scottish Office Development Department, The Welsh Office, The Department of the Environment for Northern Ireland and the Department of Transport (1996), *Highway Economics Note N°2*, in *Design Manual for Roads and Bridges*, Vol. 13, HMSO, London.

Milne, D. and D. Van Vliet (1993), 'Implementing Road User Charging in SATURN' *ITS Working Paper 410*, Institute for Transport Studies, University of Leeds, Leeds, December.

MVA (1995), *The London Congestion Charging Research Programme*, Final Report, Vol. 1: Text, Government Office for London, HMSO, London.

Newbery, D. M. (1988), 'Road Damage Externalities and Road User Charges', *Econometrica*, Vol. 56, pp. 295-316.

Newbery, D.M. (1990), 'Pricing and Congestion: Economic Principles Relevant to Pricing Roads', *Oxford Review of Economic Policy*, Vol. 6, N° 2, pp. 22-38.

Newbery, D. M. and G. Santos (1999), *Quantifying the Costs of Congestion*, End of Award Report, ESRC Grant R000222352, Department of Applied Economics, University of Cambridge, Cambridge, December.

Santos, G. (2000), 'Comparison of the Social Costs of Congestion during the Morning and Evening Peaks in Four English Towns', *Mimeo*, Department of Applied Economics, Cambridge, October.

Santos, G. and D. Newbery (2001), 'Urban Congestion Charging: Theory, Practice and Environmental Consequences', *CESifo Working Paper*, N°568, Centre for Economic Studies & Ifo Institute for Economic Research, Munich. Also available at http://www.cesIFO.de

Santos, G., Newbery, D. M. and L. Rojey (2001), 'Static Vs. Demand Sensitive Models and the Estimation of Efficient Cordon Tolls: An Exercise for Eight English Towns', *Transportation Research Board Record*, N°1747, pp. 44-50. Also published in the CD ROM of the 80th Annual Meeting of the Transportation Research, Washington DC, USA, January 7-11 '01.

Van Vliet, D. and M. Hall (1997), *SATURN 9.3 - User Manual*, The Institute for Transport Studies, University of Leeds, Leeds.