Marginal congestion cost on a dynamic expressway network

Mogens Fosgerau
Technical University of Denmark & Centre for Transport Studies, Sweden
mf@transport.dtu.dk

and

Kenneth A. Small
University of California at Irvine
ksmall@uci.edu

June 7, 2011

Keywords: congestion cost, marginal cost, dynamic congestion, queue, hypercongestion

Journal of Economic Literature code: L9, R41

Abstract

We formulate an empirical model of congestion for a series of sequential expressway links where queues may form and spill back from one link to another. Its purpose is to disentangle the dynamic effect that a marginal vehicle, on a given link and at a given time, has on the distribution of travel times experienced there and on connected links. We estimate a dynamic econometric model using unusually complete and accurate data from Danish motorways, which include speed as measured by timed license-plate matches. We use the results to estimate the marginal external cost of adding a vehicle to a link's entry flow, finding that it is highly influenced by the dynamic properties of the system of relationships between travel times and flows.

Marginal congestion cost on a dynamic expressway network

Mogens Fosgerau and Kenneth A. Small

1. Introduction

Congested road networks are receiving much attention as analysts and policy makers examine more sophisticated measures to manage traffic on existing facilities. These measures include ramp metering, express lanes, carpooling incentives, and pricing. The implications of such policies, especially express lanes and pricing, are mostly understood either from models of a single road link or from simulated networks in which road links are described by relatively simple speed-flow relationships connected, if at all, by simple queuing.

Yet the relationships spilling across links are crucial to understanding the development of highly congested systems, where queues can quickly spread and perhaps can also form spontaneously when flow approaches a saturation level. There is considerable uncertainty about the nature of flow under such conditions. It is known that on a single link, a given flow may occur at two different speeds, one relatively high and the other much lower and less stable — often called "congestion" for the former case and "hypercongestion" for the latter.[1] But there is dispute about the resulting effects on trip times over non-zero distances when conditions are rapidly changing over time and space.[2]

One common way to model severe congestion is through deterministic queuing at a bottleneck, perhaps including the spillback of queues from one link to another. Such analysis almost invariably makes the simplification that the bottleneck capacity is constant. Yet it is well known that flow tends to be unstable when it is near its maximum, and in fact capacity is often defined as the largest flow sustainable over a moderate time period rather than the maximum

[1] In much of the engineering literature, the corresponding terms are "free-flow" and "congested flow".

[2] Small and Verhoef (2007, sect 3.4.1) provide a review of these arguments. See for example the difference in opinion about the spontaneous onset of hypercongestion as a type of phase transition, reflected in Kerner and Rehborn (1997) and Daganzo, Cassidy, and Bertini (1999). Verhoef (2001) and Small and Chu (2003) argue that hypercongestion does not exist in a stable steady-state equilibrium, but rather is generated dynamically when queues form behind bottlenecks. McDonald, d'Ouville, and Liu (1999), however, claim to observe stable hypercongestion on Chicago area expressways.

possible flow achieved.[3] Furthermore, discharge rates from bottlenecks tend to rise to a temporarily high value, then fall as a queue forms, and then partially recover (Cassidy and Bertini 1999). Thus flow can exceed long-run capacity for short periods of uncertain duration, resulting in considerable stochastic variability in the travel times experienced. Furthermore, flow breakdown is persistent, leading to hysteresis (current conditions depending on the path that led to them) and requiring an inherently dynamic model to explain travel times.

Another approach, used for city street networks, is to model average flows and speeds throughout an area. Both simulation and aerial photography have suggested that such average flows and speeds can be related by an aggregate speed-flow function that has both congested and hypercongested regimes (May, Shepherd, and Bates 2000, Ardekani and Herman 1987). Small and Chu (2003) develop a dynamic aggregate model based on such a relationship that can be used to measure the marginal cost of a vehicle entering the area, but it cannot describe heterogeneity of conditions within the area. At the opposite extreme, one can model the behavior of traffic at individual signalized intersections within street networks; but this analysis becomes extremely complex when queues at one link obstruct flow on another, a situation typically requiring dynamic computer simulations with individual vehicles.

Another difficulty in modeling dynamic congestion arises in the process of empirical estimation. Such estimation requires data on the traffic flow and speed (or either of these quantities along with density) at each of many locations and times. The most common source of such data is magnetic loop detectors placed in roadways. However, the resulting data contain serious errors due to periodically non-functioning equipment and uncertain assumptions about vehicle sizes and flow homogeneity needed to convert the observed timing and spacing of axle passages into vehicle flows and speeds (Steimetz and Brownstone, 2007). Furthermore, the causal relationship between aggregate traffic flow and speed is ambiguous. The issue is that high flow may lead to reduced speed, but it may also happen that reduced speed leads to reduced flow because of increased density of vehicles impeding flow. The result is an endogeneity problem which must be dealt with econometrically.

This paper provides an empirical description of congestion formation throughout a freeway network covering a part of Denmark. We are able to solve many of the problems just

[3] See, for example, Institute of Traffic Engineers (1982), p. 471, and the *Highway Capacity Manual* published regularly by the Transportation Research Board in the United States.

described by taking advantage of an unusually detailed data set containing reliable speed measurements on each link at five-minute intervals over the entire day. Because the data are extensive, we can model congestion on these links using flexible dynamic functions. Specifically, we allow a dynamic relationship explaining travel time on a link in terms of present and past conditions on the link itself and also in terms of conditions downstream. We use instrumental variables to deal with the endogeneity of independent variables. The functional specification allows for spontaneous hypercongestion as well as hypercongestion caused by spillbacks from downstream congestion. We use the resulting model to simulate the pattern of marginal external costs associated with adding a vehicle to the traffic flow.

The results show that dynamic effects are quite important, causing perturbations in flow to persist for well over five minutes in many cases. They also show that marginal external costs arise both from the link itself, through the usual speed-flow relationship, and from the downstream link when the latter is congested. Endogeneity appears to be present and important.

The layout of the paper is as follows. Section 3 describes the data. The empirical model specification and estimation results are contained in section 4. Section 5 applies these results to calculate the marginal external cost associated with adding additional vehicles to traffic flow under some simplifying assumptions. Section 6 concludes.

2. Data

The data are collected through the period January 16 – May 8, 2007 on the freeway network in South-East Denmark.[4] The 91.1 km network links the cities of Odense in the east to Vejle in the north and Kolding in the west, as shown in Figure 1. It includes the Lillebælt Bridge, over which flows all road traffic between Copenhagen (east of Odense) and continental Denmark and Germany. Cameras are placed near each intersection, dividing the network into 15 pieces, with data recorded separately for the two directions giving observations for each of 30 one-way links. The links range from 1.7 to 11.9 kilometers in length and two to three lanes in width. Data are recorded for five-minute intervals.

[4] We are grateful to the Danish Road Directorate for providing these data.

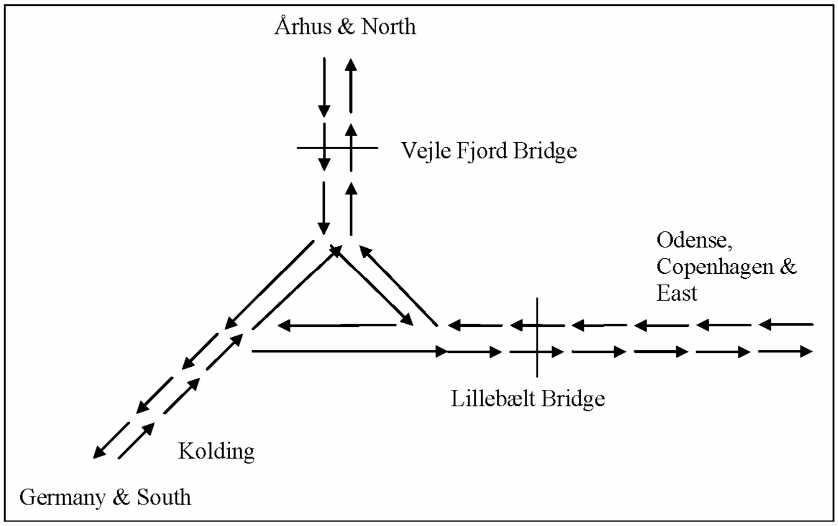

Figure 1 Network layout

We use data for all links for which we have observations on the link itself and on the first and second links upstream and downstream of it. We also require that the upstream link is a single link, in order to simplify the empirical specification and to eliminate cases where the upstream link is ambiguous. This yields 246,230 observations from nine one-way links. For every five-minute observation period, the data record the exit flows and average travel times for both light and heavy vehicles, the distinction between vehicle types being approximate as it is based on the license plate. An observation is omitted when the exit flow is less than 10 vehicles per five minutes. We compute traffic flow in passenger car equivalents (pce) using a conversion factor of 2.25 pce per truck. Travel times have been divided by distance and are expressed in minutes per kilometer, while flows are divided by number of lanes and expressed in pce per lane per minute.

We do not observe the flow within a link directly. Instead, we observe its exit flow, which we denote this F_t^n, where *n* indexes links and *t* indexes time periods. We also observe the corresponding average travel time T_t^n for the vehicles constituting this exit flow. We want to estimate the dependence of T_t^n on the flows that produced it, namely the entry flows onto link *n* over the times when those vehicles were entering. We do not observe this entry flow, but we calculate a proxy for it. Since a given link may require up to two time periods to traverse, the relevant entry flow E_t^n depends on upstream exit flow (F^{n-1}) during the current and up to two previous time periods. In the Appendix, we derive the following variable as our best approximation to the entry flow rate within the link in question:

$$E_t^n = w_0 F_t^{n-1} + w_1 F_{t-1}^{n-1} + w_2 F_{t-2}^{n-1} \tag{1}$$

with weights w_j summing to one and determined from the link length as described in the Appendix.

This calculation is inexact by nature because conditions inside the link in question may be varying over time and space. In addition, all section boundaries contain at least one entry or exit ramp, whose flows are not observed, creating an additional discrepancy between the true average flow rate and our measure of it. As discussed in the next section, we minimize the adverse impact of this measurement error by allowing travel time to depend on E_t^n through a flexible function.

We begin with some descriptive plots of the data. Figure 2 plots the observations of travel time against entry flow, with the latter (in pce per lane per minute) defined as a moving average over a one-hour period. This and later plots of the same type show, in the upper panel, a scatter plot of the data and, in the lower panel, kernel smooths giving the mean and the 95% confidence band for the mean .[5] Note the vertical scales are different in the two panels, enabling the extreme values in the raw data to be seen in the upper panel. The smoothed mean in the lower panel indicates that average travel time mostly increases slowly with flow up to a flow of about 40 pce/lane/min, after which it rises more steeply. (Its overall average in the sample is 0.57 min/km, corresponding to a speed of 105 km/h.) Although most observations are in the lower-flow region, we also have many observations of larger flows, which of course are important for measuring congestion effects.

[5] Here and later, we use a normal density kernel with bandwidth set to 5 percent of the range of the independent variable.

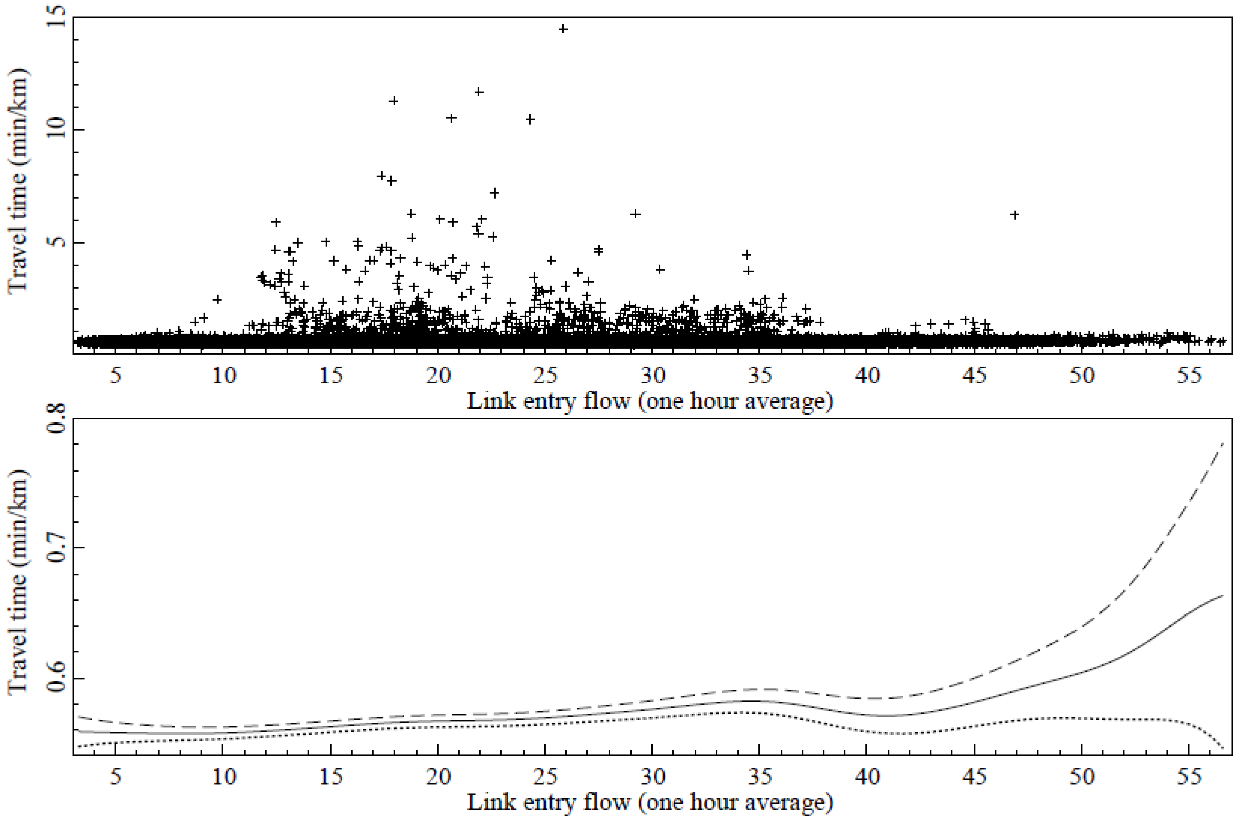

Figure 2. Link travel time (min/km) versus one-hour moving average of link entry flow (pce/lane/min); mean and 95% confidence band in lower panel

The scatter plot reveals that there is a very large dispersion of travel times: most observations are near the average but a considerable number are much larger. We believe these observations with high travel times are real and therefore we include them in the analysis; most of them occur at low entry flows, probably indicating conditions where entry flow is blocked by queues forming at bottlenecks within or downstream of the link in question.

These extremes are absent for observations above 48 pce/lane/min hourly average flow. This is probably because of the two-way causality between flow and travel time: extreme travel times are associated with queues that are likely to spill back onto the upstream link and therefore prevent high entry flows from occurring. Our empirical model captures this kind of interaction in two ways. The econometric model uses conditions on the upstream link as one of two instrumental variables to correct for the fact that flow on a given link is endogenous to travel time on the same link. And when we compute marginal cost, we explicitly consider how adding a car to one link changes conditions on the link downstream of it.

Figure 3 plots the entry flow against time of day. There are morning and afternoon peaks even though the data include both weekdays and weekends. Data are mostly missing during the hours 1:00-5:00 a.m. when there is too little traffic for reliable measurement. The lower panel shows the smoothed mean. The confidence band for the mean is present but so narrow that it is not visible.

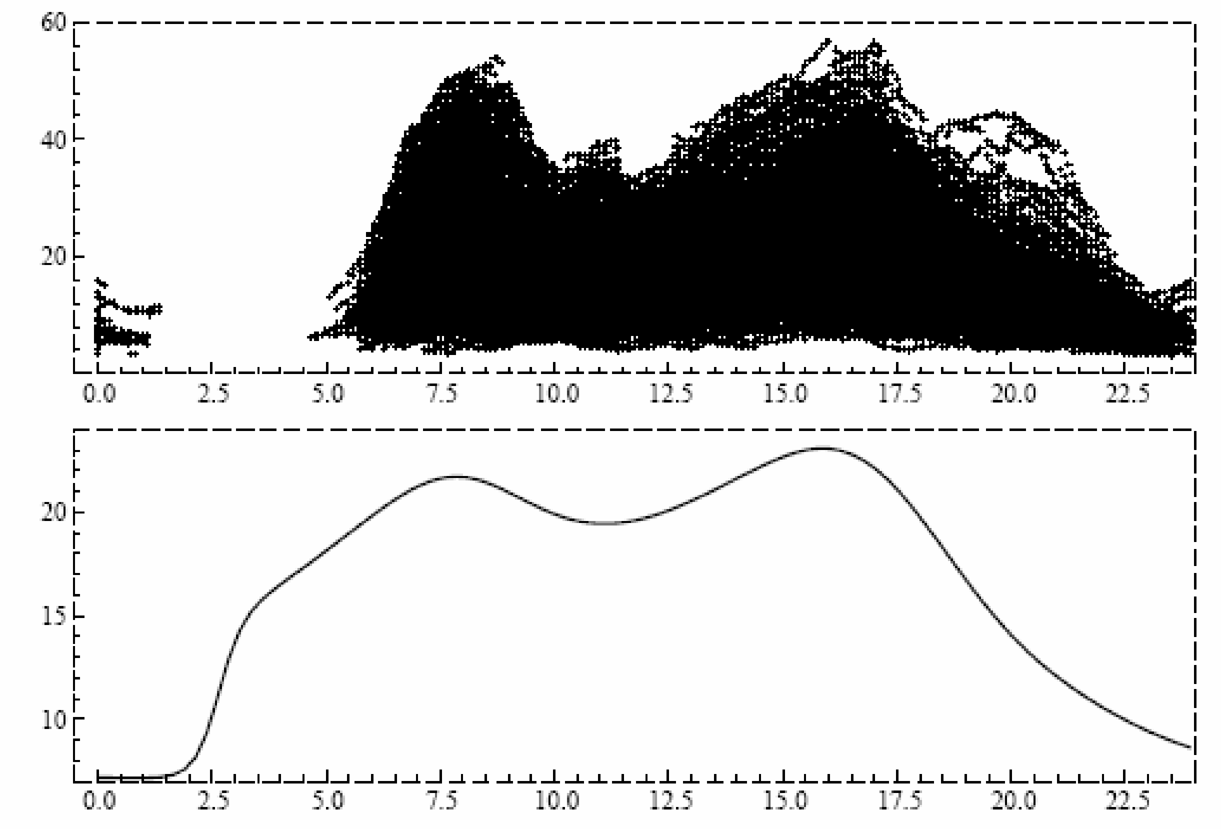

Figure 3. One-hour moving average of link entry flow (in pce/lane/min) versus time of day (hours past midnight).

3. Model specification

We now specify an empirical model to describe the dynamics and the upstream and downstream linkages that we think are most important. The model is motivated by our assessment of the most important sources of simultaneity. Specifically, we assume that current travel time on a given link depends on: (i) entry flow; (ii) lagged travel times on the same link; (iii) a proxy for queue length within the link; (iv) a measure of downstream blockage; and (v) controls for weather conditions and traffic mix, and any unmeasured physical characteristics of that particular link. Accounting for factors (i), (iii) and (iv), along with equation (1), may be regarded as an approximation to those mechanisms at the heart of the cell transmission model of Daganzo (1994), which focuses on how flows out of a road segment ("cell") into the next one downstream depend on the densities of the traffic in the two cells and on the minimum of their capacities. Unfortunately we lack the density observations that would be needed to apply the cell transmission model itself.

To implement this, we specify the logarithm of current travel time on link *n*, $\log T_t^n$, to be an additive function of five sets of variables representing the factors just mentioned. (i) First, we expect current travel time to depend on current entry flow E_t^n, just as in conventional static models using the "fundamental diagram of traffic flow" (Haight 1963, pp. 69-73). We represent this dependence as a flexible function of the natural logarithm of entry flow, $f(\log E_t^n)$. (ii) Second, we expect that current flow will be impacted by recent flow imbalances due to (unobserved) internal queuing, and thus we include two lagged values of log travel time, $\log T_{t-1}^n$ and $\log T_{t-2}^n$. (iii) Third, the existence of an internal queue is likely if entry flow exceeded exit flows in the immediate past, so we define a variable

$$Q_t^n = Max\left\{E_{t-1}^n - F_{t-1}^n, 0\right\} \tag{2}$$

where E_{t-1}^n and F_{t-1}^n are the entry and exit flows, respectively, for link *n* in the previous five-minute time period.[6] (iv) Fourth, downstream blockage will affect travel time through a term depending on downstream density D_t^{n+1}, which again we specify as a flexible function,

[6] This interpretation of the variable *Q* assumes it is multiplied by the duration of a time period, thus measured in units of passenger-car equivalents (pces). It is actually more closely related to the one-period *increase* in the queue length than the queue length itself; but it still seems a potentially useful proxy because we capture persistence in travel-time reductions another way, namely through a lagged dependent variable. It would be preferable in principle to define a variable that cumulates past imbalances between entry and exit flow; such an approach leads to numerical problems, however, due to gaps in the data and the fact that measurement errors will also cumulate. We therefore rely on the current definition as the best we can do. Variable *Q* often takes value zero, so we do not take its logarithm in the specification below.

$g(\log D_t^{n+1})$. The downstream density itself is calculated as the downstream exit flow times the downstream travel time, $D_t^{n+1} = F_t^{n+1} \cdot T_t^{n+1}$. When there are two downstream links, the average is used. (v) Finally, we include link-specific constants and two control variables, *W* and *H*, formed as follows. Control variable *W* is the logarithm of travel time experienced concurrently on the same roadway in the opposite direction (which is of the same length), averaged over the 15-minute interval centered around the current time interval; its purpose is to control for the manner in which driver behavior is affected by common elements, including weather and lighting conditions, which are both time- and location-specific. Control variable *H* is the share of heavy vehicles in current exit flow (as measured using passenger-car equivalents); its purpose is to account for the fact trucks generally travel slower than passenger cars. We also include an additive error term, ε_t^n.

To summarize, the empirical equation is:

$$\begin{aligned}\log T_t^n &= \beta_0^n + \beta_1 \log T_{t-1}^n + \beta_2 \log T_{t-2}^n + f(\log E_t^n) + g(\log D_t^{n+1}) + h(Q_t^n) \\ &\quad + \beta_W^n W_t^n + \beta_H H_t^n + \varepsilon_t^n\end{aligned} \qquad (3)$$

with E_t^n measured by (1). As discussed later, some of the right-hand-side variables, namely E_t^n, D_t^{n+1}, and Q_t^n, are endogenous.

The implied steady-state speed-density relationship is seen by substituting $T = T_{-1} = T_{-2} \equiv \overline{T}$ and ε=0 into (3), and solving with other variables held steady at values $\overline{D}^{n+1}$, $\overline{Q}^n = 0$, and $\overline{W}^n$. (Note that even if our model had included D^n as a variable, downstream density $\overline{D}^{n+1}$ could influence travel time on link *n* directly because density is measured as an average over the link. Because it is in fact is not uniform over the link, high downstream density can cause a queue that backs up into the link in question, producing different results compared to those from an increase in average density on the link itself.) The result is:

$$\log \overline{T}^n = \frac{\beta_0^n + f(\log \overline{E}^n) + g(\log \overline{D}^{n+1}) + \beta_W \overline{W}^n + \beta_H \overline{H}^n}{1 - \beta_1 - \beta_2}. \qquad (4)$$

This equation is valid provided $-1<\beta_1+\beta_2<1$, a condition that is necessary for dynamic stability and which we find true empirically in every case. Thus the effect of an exogenous shock to steady-state entry flow is determined from:

$$\frac{\partial \log \overline{T}^n}{\partial \log \overline{E}^n} = \frac{f'(\log \overline{E}^n)}{1-\beta_1-\beta_2}\,.$$

It is worth noting that by distinguishing between average flow and exit flow, our formulation solves one of the dilemmas of empirical specification of speed-flow functions. Engineering realism suggests a functional form with a maximum possible flow, such as a backward-bending speed-flow curve. But such a function cannot tell us what happens when quantity demanded exceeds capacity; furthermore, it leads to unstable and nonsensical apparent equilibria when interacted with certain demand curves (Small and Verhoef 2007, pp. 84-88). This is because flow is typically treated as a single variable, depicting both the flow that determines congestion (a supply relationship) and the quantity of travel chosen at a given level of congestion (a demand relationship). But then the backward-bending part of the speed-flow relationship makes the supply curve downward-sloping, as though one could improve conditions by adding more cars to the link. In our formulation, we can think of average flow as approximating “quantity demanded”, i.e., the amount of travel chosen by potential travelers given the conditions they face; it can exceed exit capacity without contradiction because there are entrances and exits along the link and queue lengths can change so as to absorb imbalances between link flow and exit capacity. One of the aims of our specification is to capture these factors by the dynamics in (3) and by the terms involving queuing variable *Q*.

The function *f*(·) is expected to rise slowly at low entry flows, then steeply at some value approximating the capacity of an expressway lane. After some experimentation, we find a simple piecewise linear function with one breakpoint works well. Similarly, we use one breakpoint for *g*(·) and two for *h*(·). We also experimented with cubic functions for *f, g,* and *h*; but those models do not fit as well, and also the cubic functions are overly sensitive to our numerous low-congestion observations and display regions with wrong-sign derivatives. We did, however, use the estimated cubic functions visually to choose our breakpoints for the piecewise linear specification.

We estimate link-specific fixed constant effects as well as link-specific parameters for control variable *W*. All other parameters are common across links.

Table 1 presents some descriptive statistics for the variables in the estimated equations.

Table 1. Descriptive statistics

	$\log T_{t-1}^{n}$ (log min.)	$\log D_t^{n+1}$ (log pce / lane-km)	Q_t^n (pce / lane-min)	$\log D_t^{n+1}$ (log pce / lane-km)	W_t^n (log min.)	H_t^n (percentage)
Mean	-0.5774	2.902	6.160	2.082	0.5200	0.5813
Median	-0.5913	2.961	0.000	2.081	0.5431	0.5508
Maximum	2.8420	4.226	81.41	4.588	0.9183	19.40
Minimum	-0.8519	0.831	0.000	0.089	0.1106	0.4208
Std. Dev.	0.1374	0.500	10.08	0.548	0.1428	0.3071
Skewness	5.836	-0.459	1.991	0.013	-0.364	22.4
Kurtosis	78.01	2.87	7.28	2.97	2.41	723.4

Finally, we turn to endogeneity of variables. According to the discussion in Section 2, we must regard entry flow, queue, and downstream blockage as endogenous since these variables are all affected by current congestion. We therefore use an instrumental variables (IV) estimator, which requires us to specify instrumental variables that are correlated with the endogenous variables but uncorrelated with the current residual in (3). We use the following two variables as instruments: the flow two links upstream of the current link, and the density two links downstream. The rationale for these variables is that they influence entry flow, queue size, and/or upstream density directly, but they are unlikely to be correlated with the residual in (3) because blockages seldom if ever are observed to extend across more than two links. We include also lags and some powers of these two instruments in order to gain as much power as possible in explaining the endogenous variables, while testing to avoid weak instruments as explained in the next section.

4. Estimation results

We present estimates from three models based on piecewise linear specifications of functions *f*, *g* and *h*, all using instrumental variables unless otherwise noted. The function *f* involving entry flow has a breakpoint at 40 pce/lane/min, which corresponds to the point in Figure 2 where mean travel time begins to rise and which just slightly exceeds the Danish design standard for lane capacity.[7] The function *g* involving downstream density is zero until a density of 50 pce/lane/km and linear from there. To interpret this breakpoint value, note that it corresponds to a point where downstream flow divided by downstream speed equals 50: for example, to a flow at capacity of 50 pce/lane/min and a speed of 1 km/min (60 km/h) which is roughly half free-flow speed.

The specification of the function $h(\cdot)$ varies across models. In model M1, *h* is a piecewise linear function (with two pieces) in *Q*, defined as the positive part of the sum of two lagged differences between entry and exit flow — a natural extension of (2). Model M2 replaces *Q* by its first constituents, namely the first lagged values of entry and exit flows, entered as logarithms, estimating a separate coefficient for each. Finally, model M3 omits the *Q* variable altogether.

Results are shown in Table 2. The three models are estimated in EViews by two stage least squares (TSLS). They yield an adjusted R-square of about 0.5 and a Durbin-Watson statistic close to 2, indicating little autocorrelation of the residuals. Table 2 furthermore shows the result of estimating model M3 using OLS.

All three models portray stable and statistically significant dynamics. The coefficients for first and second lags of travel time are positive, very significant, and sum to less than one (about 0.7). These values imply that the remaining coefficients should be multiplied by about $1/(1-0.7)\approx3.3$ to get the values that apply when the model is solved in steady state, as shown in equation (4).

[7] The Danish design standard states an ideal capacity of 2300 pce/lane/h, or 38.3 pce/lane/min (www.vejregler.dk).

Table 2. Estimation results

Dependent variable: natural logarithm of travel time per km (min/km)								
Model:	M1		M2		M3		M3 OLS	
Variable	Coeff.	t-Stat.	Coeff.	t-Stat.	Coeff.	t-Stat.	Coeff.	t-Stat.
Const.	-0.261	-34.2	-0.253	-32.0	-0.264	-33.6	-	-
T_{-1}	0.379	75.5	0.375	74.8	0.381	76.8	0.395	108.4
T_{-2}	0.309	62.6	0.308	61.2	0.310	63.5	0.323	89.8
lnE	0.007	3.8	0.075	5.4	0.008	4.2	0.005	5.4
$(\ln E-\ln(40))*1_{\{E>40\}}$	0.232	4.0	0.178	3.3	0.144	2.7	0.054	4.0
$(Q-20)*1_{\{Q>20\}}$	-0.001	-4.0						
$(Q-40)*1_{\{(Q>40\}}$	-0.001	-1.7						
$\ln E_{-1}$			-0.070	-5.6				
$\ln F_{-1}$			-0.001	-0.7				
$(\ln D^{n+1}-\ln(50))*1_{\{D^{n+1}>50\}}$	3.583	5.0	4.449	6.2	3.679	5.1	0.849	13.5
H	0.045	13.5	0.042	12.5	0.044	13.6	0.037	13.4
link-specific constants	yes		yes		yes		yes	
link-specific const's * *W*	yes		yes		yes		yes	
Number of observations	66,902		67,352		67,777		67,777	
Adjusted R-squared	0.503		0.495		0.509		0.524	
Sum squared resid (SSR)	645.214		663.418		652.246		632.680	
Durbin-Watson stat	2.083		2.049		2.089		2.160	
Second-stage SSR	627.699		631.086		633.877		n.a.	

Note: The symbol 1_A denotes the indicator function for event *A.*

Other control variables are also stable across models. The coefficient for *H,* the share of heavy vehicles in the exit flow, indicates that the travel time of heavy vehicles is 13–15 percent larger than for light vehicles in the same traffic stream.[8] With nine links included, there are nine link-specific effects of control variable *W*. The latter control variables almost all have

[8] That is, for given values of other right-hand-side variables, the travel times for truck and car, T_T and T_C, are related by $T_T/T_C = \exp[\beta_H/(1-\beta_1 -\beta_2)] \approx 1.13$–$1.15$. By comparison, the ratio of speed limits for cars to trucks is 1.38–1.62;.but most trucks go faster than the speed limit. We tried including interactions between the share of heavy vehicles and the functions for entry flow and downstream density, but these interactions were jointly insignificant. A more complicated alternative would be to develop a model with travel time for cars and trucks as separate dependent variables.

statistically significant effects, typically in the range of 0.02–0.15, indicating that travel time on the opposing link is mildly correlated with travel time on the link in question.[9]

Turning to the variables of main interest, consider first the role of our queuing proxy, *Q*, in explaining travel time. Model M1 represents the effect of *Q* as two linear pieces, one for *Q* between 20 and 40 and one for *Q* above 40. We expect a positive relationship because an internal queue should cause delay; however, the estimated coefficients are both negative. In model M2, we replace *Q* by the entry and exit flow of the last period; we see that the lagged exit flow becomes insignificant and that the lagged entry flow receives a large negative parameter, while the parameter for current entry flow (already positive in model M1) becomes much larger . We conclude that this variable does a poor job of capturing the effect of internal queuing, not surprisingly given the discussion in section 2.1. In particular, the entry flow is not measured but is approximated in (1) as a weighted average of past exit flows from the upstream link; and we have no information on how many vehicles enter and exit the freeway along the way. There is also the possibility that the effect of internal queuing is just not very strong in our dataset.

In model M3, we therefore discard the internal queuing variable. The results are reassuringly similar to model M1 except that the effect of downstream density is greater. We therefore consider this our most reliable model. Model M3 shows steady-state elasticities exceeding the one-period elasticities by a factor of $(1\text{-}0.381\text{-}0.310)^{-1}$=3.2. Entry flow has a significant positive effect on travel time, with a steady-state elasticity about 0.008*3.2≈ 0.026 at entry flows less than 40 pce/lane/min and a much larger elasticity of 0.152*3.2≈0.49 for larger flows. The coefficient for the downstream density implies a large steady-state elasticity of 3.679*3.2=11.9 when density is greater than the breakpoint. Our results thus confirm that queue spillbacks can be an important contribution to congestion, as has long been assumed throughout the engineering literature (e.g. May 1990).

Just how high a degree of statistical significance should we expect from this model? We note that the effective sample size for the congestion variables is much smaller than the full sample size, because most observations show no congestion (Fig. 1). Therefore, the moderate asymptotic t-statistics we find for the associated parameters, typically between four and six,

[9] In addition the estimation procedure effectively estimates a fixed effect model, but without explicitly estimating the fixed-effect coefficients, by subtracting from each independent variable its mean value (across time) for a given link.

seem satisfactory. But for the same reason, our estimation routine breaks down (e.g fails to converge) when we try to estimate the model while allowing for more detailed dynamics such as more lagged values of travel time, first-order correlation, or different behavior when congestion is growing versus when it is declining. Thus, we cannot rule out more complex dynamics, or the possibility that unmeasured factors persisting across time could be explaining some of the dynamic behavior that we attribute to traffic interactions. It appears we have reached the limits of this data set to distinguish among the various possibilities.

The last columns of Table 2 show for comparison the results of estimating M3 using OLS; that is, without taking endogeneity of entry flow and downstream density into account. While most parameter estimates are largely unaffected, those corresponding to the endogenous variables change markedly as is expected when endogeneity is present and important.

The instruments were chosen by removing instruments, one by one, from the full list of potential instruments described earlier, until the Sargan test indicated acceptance of over-identifying restrictions . The corresponding regression of residuals against instruments yielded a significance level of 0.59, indicating that residuals are not seriously correlated with instruments. This is in agreement with the maintained hypothesis of model M3.

To check the strength of the instruments, we furthermore carried out the first-stage regressions in a separate procedure, regressing the endogenous variables on all exogenous variables and testing the joint significance of the instruments (i.e., the significance of the exogenous variables that do not enter directly in the model). For each of the endogenous variables we found that the instruments were very significant in explaining the endogenous variables.[10] The first-stage estimates are shown in the Appendix.

5. Illustrative calculation of marginal external costs

This section presents a calculation of marginal external cost (*mec*), in time units. We caution that because we observe a situation where no congestion tolls are being collected, there is no reason to identify the resulting marginal external costs with an optimal toll schedule. Nevertheless, knowledge of the pattern of the *mec* does provide some preliminary guidance

[10] The F-statistics are 1693.7 for variable (lnE); 222.1 for variable $[(\ln E-\ln(40)]*1_{\{E>40\}})$; and 11.5 for variable $((\ln D^{n+1}-\ln(50))*1_{\{Dn+1>50\}})$. Corresponding significance levels in the F-distribution are virtually zero.

as to toll design, and also tells us the social value of other policy interventions that would reduce entering traffic to a given link at a given point in time.

Calculation of the *mec* is difficult, since it must take into account that both nonlinearity and dynamics are present. The nonlinearity implies that simplifications using averages are not available. In particular, the *mec* depends in a nonlinear way on entry flow and downstream density as well as on the residual. Moreover, the fact that the model is dynamic means that the effect of a marginal vehicle depends on the entry flow, downstream density, and residuals not only at current but also at later times.

One appealing possibility which we ultimately decided against would be to use the dataset and the fitted model to generate a sequence of residuals in the equation explaining travel times. These residuals could be used to generate simulations replicating the observed data; the simulations could then be repeated after adding a vehicle at a specific point in time, thereby generating one *mec* value. If this process were repeated for every time point in the data, the results could be averaged to find the expected *mec* as a function of the time of day. But such a procedure would encounter the problem that there are gaps in the data, causing bias when congestion that reaches into gaps is omitted.

For this reason we have chosen an alternative approach to computing *mec*. The approach employs some simplifications and therefore the results must be regarded as illustrative of the model.

We begin by observing that with average flow *E* and travel time *T*, the internal cost (ignoring monetary costs) is *T*, total cost to all users is *TE,* and

$$mec = \frac{\partial(TE)}{\partial E} - T = \frac{\partial T}{\partial E} E \,. \tag{5}$$

For ease of computation and interpretation, we consider the case where *E* and *T* initially take steady-state values, operating according to our preferred Model 3. The conceptual experiment in (5), however, incorporates the dynamics of the system since it involves a perturbation from this steady state. Furthermore, we consider the case where the upstream link starts with the same initial conditions as the current link. (This does not preclude the *downstream* link from being different, for example by having a smaller capacity and thus creating a bottleneck.) Writing out equation (4) for our preferred empirical model and omitting terms that are unaffected by adding an additional car to the current link, we then have:

$$\log T = \gamma_1 \log E + \gamma_2 \log\left(\frac{E}{40}\right) \cdot 1_{\{E^n > 40\}} + \gamma_3 \log\left(\frac{D^{+1}}{50}\right) \cdot 1_{\{D^{+1} > 50\}} \tag{6}$$

where γ denotes the appropriate coefficient from Table 2 divided by (1-β_1-β_2). In (6), we have suppressed the link-specific notation *n* as well as the bar indicating steady state.

Equation (6) depicts how a given exogenous increment to flow will affect travel time in two ways. First, it directly affects the current link through the terms involving γ_1 and γ_2. Second, it indirectly affects the link upstream of it through the term involving γ_3 by adding to the density *D* of the current link, as can be seen by rewriting (6) for that upstream link:

$$\log T^{-1} = \gamma_1 \log E^{-1} + \gamma_2 \log\left(\frac{E^{-1}}{40}\right) \cdot 1_{\left\{E^{-1^n} > 40\right\}} + \gamma_3 \log\left(\frac{D}{50}\right) \cdot 1_{\{D > 50\}}$$

where superscript -1 denotes the upstream link. The derivative $\partial T/\partial E$ in (5) is the sum of these two effects, and so *mec* can be written as the sum of two components:[11]

$$mec_E = \left(\gamma_1 + \gamma_2 \cdot 1_{\{E>40\}}\right) T\,. \tag{7}$$

and

$$\begin{aligned} mec_D &= \frac{\partial T^{-1}}{\partial D} \cdot \frac{\partial D}{\partial E} \cdot E \\ &= \gamma_3 \cdot 1_{\{D>50\}} \cdot \frac{TE}{D} \cdot \frac{\partial D}{\partial E} \end{aligned}\,. \tag{8}$$

Applying (5) and the identity *D*=*TE*, (8) simplifies to

[11] This decomposition is analogous to those of Yang and Huang (1998), who also explicitly consider upstream and downstream links, and to that of Mun (1999), who divides the link being analyzed into a queued portion and a portion subject to normal congestion. Going further afield, *viz.* to the second upstream link, would most likely lead to negligible additions to *mec*.

$$mec_D = \gamma_3 \left(mec_E + T \right) \cdot 1_{\{D > 50\}} \quad (9)$$

We compute both components of *mec* from (7) and (9) for every observation in our sample, thus depicting for each value of *n* and *t* what the *mec* would be if the flow, travel time, and downstream density for that observation were maintained for several periods. An advantage of using (9) in this way is that the formula uses realisations of *T*, including its error term. This matters for the result as the dependency of *T* on the error term in (3) is nonlinear and so the distribution of error terms is important. The present formula preserves this information in a way that is easy to handle.

Figure 4 presents a scatter plot of the *mec* against the entry flow, the latter expressed as an hourly average. Each data point on the scatter corresponds to a five minute interval on a section in the network. The scatter in the vertical direction is therefore due to both variation in entry flow within this one-hour average and to the presence of a random term in (3), which is therefore incorporated into the value of travel time *T* as it appears in equations (7) and (9). The discontinuities caused by the indicators for large flow and density in (7) and (9) are visible as vertical gaps between clouds of *mec* values.[12,13] Furthermore, because the distribution of travel time *T* for any given initial flow is positively skewed, so is the distribution of marginal external cost. Again, the lower panel shows the smoothed mean of the *mec*, as well as the 2.5 and 97.5 percentiles. The smoothed mean may be considered an estimate of the expected *mec* conditional on the initial average flow.

The mean *mec* is initially small and rises slowly until a flow of about 30 pce/lane/min. At this point the *mec* of a vehicle is about 0.15 min/km, which corresponds to about 25 percent of the average travel time (and a much larger proportion of the private congestion delay). From this point, the *mec* rises more steeply and at a flow of 40, the nominal capacity of a lane, *mec* has reached 0.25 min/km or a little more than half the average travel time. At 50 pce/lane/min, *mec*

[12] Note the logarithmic scale used for the upper panel; the two gaps occur at very different values of *mec*.

[13] If we had used differentiable rather than piecewise linear specifications for functions $f(\cdot)$ and $g(\cdot)$ in (3), these “clouds” would not be detached as they are here. However, functional form assumptions are hard to verify and the shapes of f and g are not well identified from the data. So we believe little would be gained from using differentiable functions, especially since the gaps are eliminated when we present curves depicting average.

has risen to about 0.55 min/km which is more than the (increased) travel time. This result confirms the view, expressed in many economic models of congestion, that external cost rises slowly at first, then rapidly as the entry flow approaches and then exceeds the capacity of a highway. This rising *mec* confirms common perceptions and is quite important for the welfare effects of pricing policies.

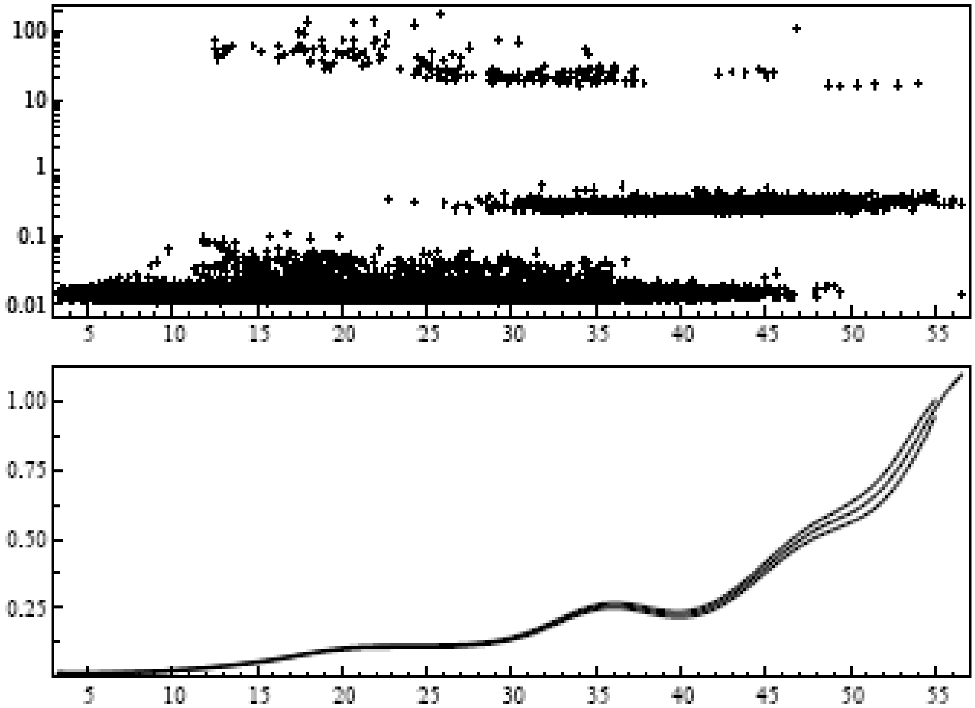

Figure 4. *mec* versus hourly average flow

Figures 5 and 6 present the two components of the *mec* separately. It seems the component reflecting entry flow congestion, mec_E, is extremely variable and its average dominates until the component reflecting downstream congestion, mec_D, becomes the larger.

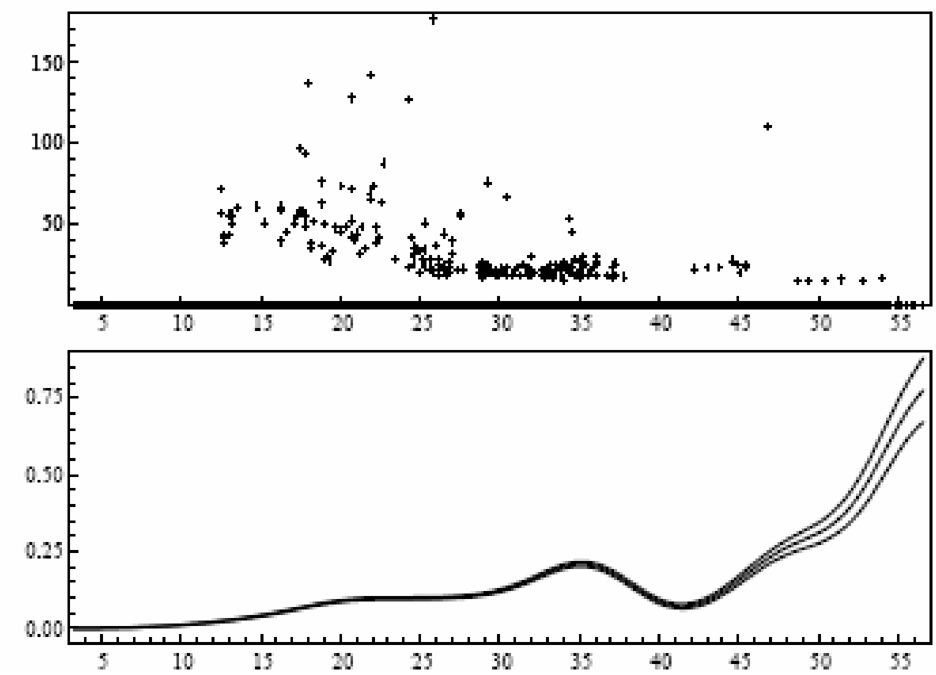

Figure 5. The contribution of entry flow to the *mec*: *mec*$_E$ versus hourly average flow

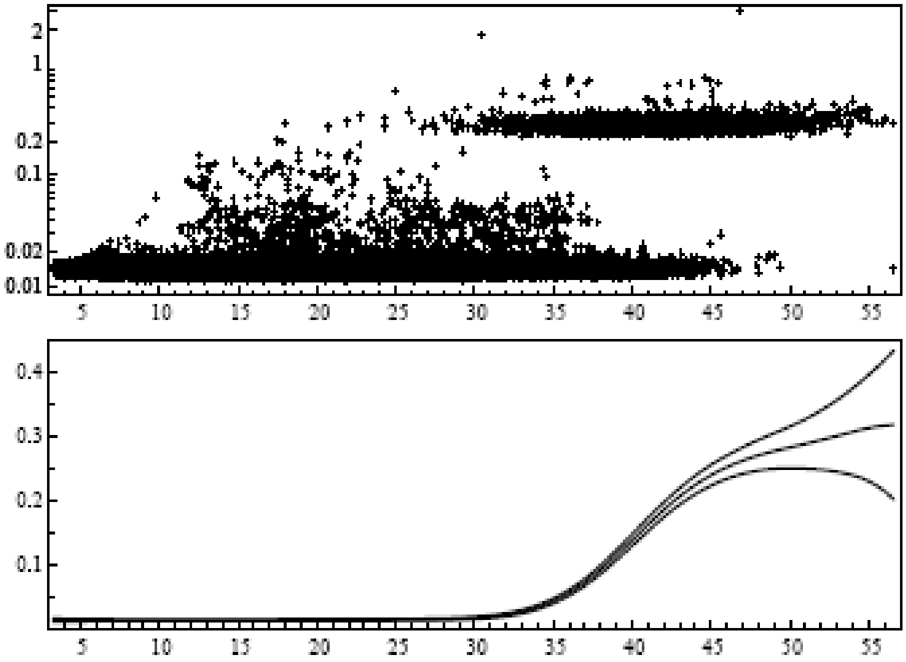

Figure 6. The contribution of downstream density to the *mec*: *mec*$_D$ versus hourly average flow

Finally, Figure 7 shows the *mec* against the time of day. Its average follows the peaks in traffic but seems to be highest at around 2:30 p.m. Evidently, on this network the lower-flow situations are more common than the higher-flow situations even at the peaks, causing the average *mec* to be well below the values shown in the cloud of calculated points at high average flows. Of course, there are many individual data points where *mec* is much higher than this, a reminder that marginal external cost can vary a lot due to randomness in conditions.

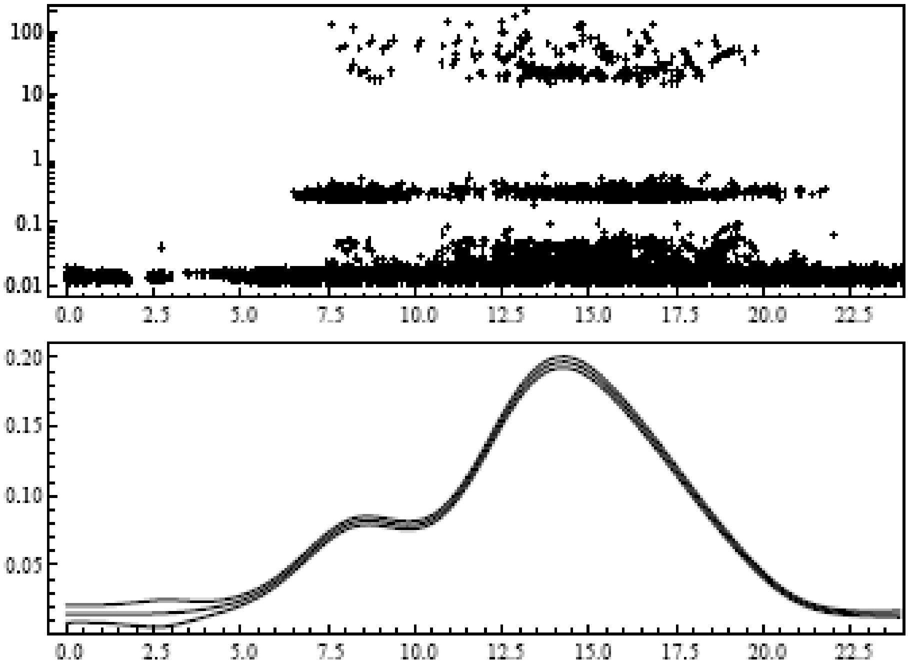

Figure 7. *mec* versus time of day

We note also that average *mec* is much higher during the afternoon than the morning peak, despite only a small difference in traffic volumes as shown in Figure 3; and that it peaks earlier than traffic volumes. This is because most episodes of high travel times occur during the early afternoon in our data, and because of the highly nonlinear nature of equation (3) they result in high values of *mec* which pull up the average *mec* during those hours.

6. Caveats and Conclusions

Our results confirm the potential importance of congestion dynamics for understanding how an exogenous change in traffic entering a network affects the resulting travel costs. We find that when the system is initially in a steady state, the ultimate impact of adding a vehicle is more than three times the immediate impact, due to the persistence of congestion as indicated by the estimated dynamic properties of our model. These dynamic effects have rarely been measured, and we feel sure there is room for much more empirical work measuring them. In addition, there are of course many more initial conditions that could be analyzed to see where marginal external costs are high and where they are low.

We noted in Section 4 that the estimated coefficients of the lagged travel times are sensitive to the specification of autocorrelation in the error terms. If those lagged travel times are actually serving as proxies for autocorrelation resulting from persistent influences not captured by our specification, the time-series correlations in the data would not be explained by congestion dynamics and thus would not raise the external costs in the manner calculated here. Therefore, further research providing greater precision in measuring autocorrelation would help refine the quantitative magnitudes of these external costs.

We have also provided a way to handle simultaneity of speed and traffic flow in a statistical model, using observations from other links in the same network as instruments for flow in an equation describing travel time as a function of flow. Our results using these instruments indicate a positive and increasing marginal external cost in accordance with the a priori expectation, in contrast to conventional models which have great difficulty dealing with regions of data where travel time and flow are negatively correlated. But we acknowledge the sensitivity of these results with respect to specification of the time-series properties of the residuals and therefore the need for better data in order to confirm our results.

Another feature of our model is the effect of downstream congestion on the travel time on the current link. This effect is found to be empirically significant and important for the marginal external cost.

References

Ardekani, Siamak, and Robert Herman (1987), "Urban Network-Wide Traffic Variables and Their Relations," *Transportation Science*, 21: 1-16.

Cassidy, Michael J. and Robert L. Bertini (1999), “Some traffic features at freeway bottlenecks,” *Transportation Research Part B*, 33: 25-42.

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

Daganzo, Carlos F., Michael J. Cassidy and Robert L. Bertini (1999), "Possible explanations of phase transitions in highway traffic," *Transportation Research Part A*, 33: 365-379.

Haight, Frank (1963), *Mathematical Theories of Traffic Flow*, New York: Academic Press.

Institute of Traffic Engineers (1982), *Transportation and Traffic Engineering Handbook*, Englewood Cliffs, New Jersey: Prentice-Hall.

Kerner, B.S. and H. Rehborn (1997), "Experimental properties of phase transitions in traffic flow," *Physical Review Letters* 79: 4030-4033.

May, Adolf D. (1990), *Traffic Flow Fundamentals*, Upper Saddle River, NJ: Prentice-Hall.

May, Anthony D., S.P. Shepherd, and J.J. Bates (2000), "Supply Curves for Urban Road Networks," *Journal of Transport Economics and Policy,* 34: 261-290.

McDonald, John.F., Edward d’Ouville, and Louis Nan Liu (1999) *Economics of Urban Highway Congestion and Pricing*, Kluwer.

Mun, Se-il (1999) "Peak-load pricing of a bottleneck with traffic jam," *Journal of Urban Economics* 46: 323-349.

Small, Kenneth A. and Xuehao Chu (2003), "Hypercongestion," *Journal of Transport Economics and Policy* 37: 319-352.

Small, Kenneth A. and Erik T. Verhoef (2007), *The Economics of Urban Transportation*, London and New York: Routledge.

Steimetz, Seiji S.C., and David Brownstone (2007), “Estimating commuters’ ‘value of time’ with noisy data: a multiple imputation approach,” *Transportation Research Part B*, 39: 865-889.

Verhoef, Erik T. (2001), “An integrated dynamic model of road traffic congestion based on simple car-following theory: Exploring hypercongestion,” *Journal of Urban Economics* 49: 505-542.

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

Notation

F = exit flow from a link (pce/min per lane)
E = average flow over a link (pce/min per lane)
T = travel time (minutes/km)
D = density (pce/lane-km)
Q = proxy for size of internal queue on link (pce/lane)
W = travel time on control section (proxy for weather, etc.)
H = the share of heavy vehicles in the exit flow
L = length of link (km)
t = time period (integer)
n = section (larger numbers are downstream)

Appendix A: Approximating entering flow from observed upstream flows

We observe the average time T_t^n required to traverse link *n* by vehicles that exit the link during a fixed five-minute interval *t*, which here we take to begin at time 0 and end at time 5. We want to model T_t^n as a function of the entry-flow rate E_t^n which was encountered by those same vehicles when they entered that link. We do not observe this entry flow directly, but can approximate it from what we know of the exit flow of the link immediately upstream (link *n*-1). Specifically, we want to know that exit flow, F^{n-1}, during the time period(s) when the vehicles in question would have been at that location. Knowing the length of the link, L^n, and the average speed on it, we can compute which time intervals are relevant. In order not to introduce endogeneity into the calculation, we use the average speed on the entire network, S* (expressed in km/min).

None of our sections take longer than two time periods (ten minutes) to traverse. Therefore there are up to two previous time periods whose upstream entry flows are relevant: just one in the case of short sections (those that take less than five minutes to traverse), and two in the case of longer sections. Thus our entry flow variable will be a weighted average of the observed upstream flows in the current and up to two previous time period, with the weights equal to the proportions of vehicles that could be expected to have been observed during those time periods:

$$E_t^n = w_t F_t^{n-1} + w_{t-1} F_{t-1}^{n-1} + w_{t-2} F_{t-2}^{n-1}.$$

Consider the vehicles exiting link *n* during the five-minute time interval *t*, i.e. between time 0 and time 5. For a link with length $L \leq 5S^*$, all the vehicles exiting before time $L/(5S^*)$ entered the section during interval *t*-1, while the rest entered during interval *t*; so $w_{t-1}=L/(5S^*)$ and $w_t=1-L/(5S^*)$. For a longer link, all the vehicles exiting before time $[-5+ L/(5S^*)]$ entered during interval *t*-2, the rest during interval *t*-1; so $w_{t-2}=L/(5S^*)-1$ and $w_{t-1}=1-L/(5S^*)$. We can summarize both cases as follows:

$$\begin{aligned} w_t &= Max\{0,\ [1 - L/(5S^*)]\} \\ w_{t-2} &= Max\{0,\ [L/(5S^*) - 1]\} \\ w_{t-1} &= 1 - w_t - w_{t-2} \end{aligned}$$

Appendix B: First-stage regression for Model M3

Table B1gives results for the first-stage regressions, *i.e.* those explaining the endogenous right-hand-side variables in (3), for the two-stage least squares estimation of Model M3 of Table 2.

Table B1. First-stage regression results for Model M3

Dependent variable:	lnE		$(lnE-ln(40))*1_{\{E>40\}}$		$(lnD^{n+1}-ln(50))*1_{\{Dn+1>50\}}$	
Independent variable	Coeff.	t-Stat.	Coeff.	t-Stat.	Coeff.	t-Stat.
C	1.316	93.8	-0.026	-18.4	0.007	22.1
T_{-1}	0.002	0.2	0.008	7.8	0.004	19.5
T_{-2}	-0.047	-4.7	0.000	-0.5	0.004	19.7
$(lnF^{n-2}-ln10)*1\{F^{n-2}>10\}$	0.263	38.9	-0.006	-8.6	0.000	-0.8
$(lnF^{n-2}-ln20)*1\{F^{n-2}>20\}$	0.009	0.6	0.023	15.9	0.000	0.9
$(lnF^{n-2}-ln30)*1\{F^{n-2}>30\}$	0.050	1.5	0.070	21.0	-0.001	-1.6
$(lnF^{n-2}-ln40)*1\{F^{n-2}>40\}$	-0.130	-1.6	0.035	4.4	0.001	0.7
$(lnF^{n-2}-ln50)*1\{F^{n-2}>50\}$	-0.241	-1.2	-0.188	-9.4	-0.002	-0.4
$ln(F^{n-2})_{-2}$	0.356	82.4	0.008	17.7	0.000	-0.8
$ln(D^{n+2})_{-1}$	0.128	23.8	0.001	1.9	0.000	-0.3
$ln(D^{n+2})_{-2}$	0.086	16.3	0.002	3.7	0.000	0.9
$(ln(D^{n+2})-ln15)*1\{D^{n+2}>15\}$	0.026	2.5	0.011	10.1	0.000	0.2
$(ln(D^{n+2})-ln30)*1\{D^{n+2}>30\}$	-0.416	-9.7	-0.050	-11.7	0.016	16.6
$(ln(D^{n+2})-ln45)*1\{D^{n+2}>45\}$	0.464	7.2	0.044	6.9	-0.019	-13.3
H	-0.037	-4.8	0.004	4.6	-0.002	-12.1
link-specific constants	yes		yes		yes	
link-specific const's * *W*	yes		yes		yes	
Number of observations	67,777		67,777		67,777	
Adjusted R-squared	0.702		0.162		0.043	