Marginal congestion cost on a dynamic network with queue spillbacks

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)

December 15, 2008
Preliminary Draft – Please do not cite or quote

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 network 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 model, based on an unusually complete and accurate dataset from Danish motorways. Each data point contains information on the vehicle flow on a link during a five-minute interval, along with the average speed experienced by those vehicles 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, as it is influenced by conditions on that link and on its downstream neighbor.

Marginal congestion cost on a dynamic network with queue spillbacks

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. We shall use the terminology, common in economics, of "congestion" for the former case and "hypercongestion" for the latter.[1] But the exact process of transition from one to the other is much debated and seems to depend critically on how one link interacts with another.[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 invariable makes the simplification that the bottleneck capacity is constant. Yet Cassidy and Bertini (1999) find empirical evidence that discharge rates from bottlenecks fall after queue formation and then partially recover. They find the higher pre-collapse flow to be unstable and they "view the [lower] long-run queue discharge flow as the bottleneck capacity" (p. 40). Thus in their view, flow can exceed capacity for short periods of uncertain duration, resulting in

[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.

considerable stochastic variability in the travel times experienced and the marginal effects of an additional vehicle.

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, 2005).

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 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. The functional specification allows for both spontaneous hypercongestion and for hypercongestion caused by spillbacks from downstream congestion, thus allowing the data to determine the relative importance of these two causes of severe congestion. We use the resulting model to simulate the pattern of marginal external costs association with adding a vehicle to the traffic flow at various demand levels.

The results show that dynamic effects are quite important, causing perturbations in flow to persist for several of our five-minute time intervals. 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 it is congested. However, our results are quite sensitive to details of model specification, leading us to suspect that our approximate solution to a full reduced-form model of travel flow by link does not capture all the interactions that occur under heavy congestion.

The layout of the paper is as follows. The model is specified in Section 2, while Section 3 provides description of 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. Section 6 concludes.

2. Model Specification

The links on our network, indexed by *n*, are defined as sections of roadway between two intersections. The time periods, indexed by *t*, are 5 minutes in duration.

2.1 Structural model of flows, queues, and congestion delay

Let T_t^n be the link travel time (in minutes per kilometer) observed for vehicles exiting link *n*, with physical length L^n km, during time interval *t*. Vehicles exit the link at rate F_t^n. These are the quantities on which we have direct measurements.

We assume exit flow F_t^n is the smaller of the potential flow reaching the end of the section and the capacity of the section, the latter being reduced by blockages from the downstream link. This potential exit flow rate is equal to the entering rate E_t^n to the link plus the discharge over time interval Δt of any accumulated internal queue, q_t^n. Downstream blockage depends in some unknown way on downstream density D_t^{n+1} (measured in vehicles per lane-kilometer).Thus:

$$F_t^n = Min\left\{\left[E_t^n + \left(q_t^n / \Delta t\right)\right], g_1(D_t^{n+1})\right\} \quad (1)$$

where $g_1(\cdot)$ is a strongly nonlinearly decreasing function. The size of the internal queue is an accumulation of past excesses of entry flows over exit flows:

$$q_t^n = Max\left\{\sum_{t'=t_0^n}^{t-1} \Delta q_{t'}^n, 0\right\}; \qquad \Delta q_{t'}^n = \left(E_{t'}^n - F_{t'}^n\right)\cdot \Delta t \tag{2}$$

where t_0^n is the most recent time period t' for which $q_{t'}^n = 0$.

Travel time follows a speed-density relationship:

$$T_t^n = h_1(D_t^n) \tag{3}$$

where by the usual traffic-flow equality, density D is given by flow F divided by speed S=1/T:

$$D_t^n = T_t^n \cdot F_t^n . \tag{4}$$

Finally, we approximate entry flow during interval t based on what we know of the exit flows from the upstream link at previous times. (This will be inexact because we lack data on entry and exit ramps.) Due to the lengths of our sections and the five-minute duration of our time interval, we need the upstream flow for the current and up to two previous time periods. The result is

$$E_t^n = w_0 F_t^{n-1} + w_1 F_{t-1}^{n-1} + w_2 F_{t-2}^{n-1} \tag{5}$$

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

Equations (1)–(5) form a simultaneous system in various flows and travel times. We would like to solve them for the flows and travel times as functions of other variables. These endogenous variables affect each other in several highly nonlinear and interconnected ways. First, T appears on both sides of (3), since it is part of definition (4) of density. Second, current values are highly nonlinear functions of lagged values through queue formation as described in (2). Third, values for section n are functions of values for downstream sections through the term $h(\cdot)$ in (1) (blockage from downstream congestion); and they depend on upstream sections through the last term on the right-hand side of (5) (entry to section n depends on exit from section n-1). Of course, the same equations apply to these upstream and downstream sections, so that congestion effects on a given section can propagate in both directions.

For these reasons, we find it intractable to estimate equations (1)–(5) structurally or to solve them explicitly for the endogenous variables. Instead, we suggest the following heuristic approximation of a solution for travel time T_t^n. It is motivated by our assessment of the most

important sources of simultaneity. First, the solution will imply a strong dependence of current travel time on entry flow, which we represent as a flexible function $f(E_t^n)$. Second, the impact of recent flow imbalances via current queue length, q_t^n, will be closely related to recent past values of travel times; we therefore approximate it by including in our reduced-form equation two lagged values of travel time, T_{t-1}^n and T_{t-2}^n. It is important for our later simulations to recognize that these lagged travel times represent congestion dynamics and therefore play a significant role in the response of the system to any perturbation of entry flow. Third, the impact of the queue will depend strongly on recent past flow differences, which we proxy in some specifications by including the variable:

$$Q_t^n = Max\left\{E_{t-1}^n - F_{t-1}^n, 0\right\}. \tag{6}$$

Fourth, upstream blockage will affect travel time through the term $g_1(\cdot)$ in (1); we approximate this effect by including a flexible function $g(D_t^{n+1})$ in the reduced-form equation for T_t^n. Finally, we include link-specific constants and two control variables, *W* and *H*, as explained in Section 4.1.

In order to ensure that error terms can take infinite values without introducing contradictions, we represent most variables by logarithms, except for *Q* which is often zero. The result is the following empirical equation:

$$\begin{aligned}\log T_t^n &= \beta_0^n + \beta_1 \log T_{t-1}^n + \beta_2 \log T_{t-2}^n + f(\log E_t^n) + g(\log D_t^{n+1}) + h(Q_t^n) \\ &+ \beta_W^n W_t^n + \beta_H H_t^n + \varepsilon_t^n\end{aligned} \tag{7}$$

The implied steady-state speed-density relationship is seen by substituting $T = T_{-1} = T_{-2} \equiv \overline{T}$ and ε=0 into (7), and solving with other variables held steady at values $\overline{D}^{n+1}$, $\overline{Q}^n$, and $\overline{W}^n$. The result is:

$$\log \overline{T}^n = \frac{\beta_0^n + f(\overline{E}^n) + g(\overline{D}^{n+1}) + h(\overline{Q}^n) + \beta_W \overline{W}^n + \beta_H \overline{H}^n}{1 - \beta_1 - \beta_2} \tag{8}$$

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 a change in steady-state entry flow is:

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

2.2 Functional Forms

The function $f(\cdot)$ 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. The same is true for $g(\cdot)$. We also experimented with cubic functions for f and g, but they do not fit as well and are less satisfactory theoretically because they contain regions with wrong-sign derivatives.

It is worth noting that by distinguishing between entry 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. 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 entry flow as quantity demanded; 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 entry flow and exit capacity. We hope that the net effect of these factors is captured by the dynamics in (7) and of the terms involving queuing variable Q.

3. Data

The data are collected through the period January 16 – May 8, 2007 on the freeway network in South-East Denmark.[3] The 91.1 km network forms a triangle with corners linking 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

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

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.

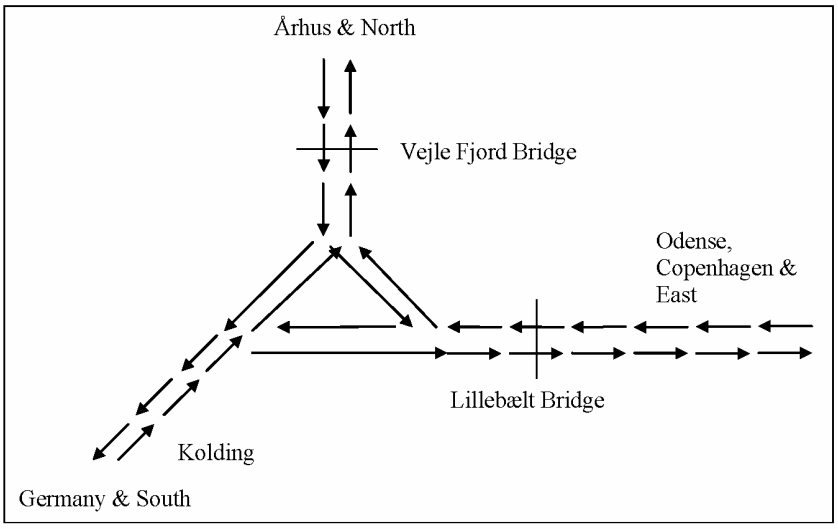

Figure 1 Network layout

We use data for all links that are joined upstream to just one other link, and for which we have observations on the link itself and on the first and second links upstream and downstream of it. 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.

Figure 2 plots the observations of travel time against flow, with the latter averaged 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, a kernel smooth including the mean and 95 percent pointwise confidence band. Using a normal density kernel, the bandwidth for the smooth here and later has been set to 5 percent of the range of the independent variable, which in Figure 2 is flow. The smoothed mean indicates that average travel time increases slowly with flow up to a flow of about 35 pce/lane/min, which is roughly the design capacity of a freeway lane; then it is almost

flat up to 45 pce/lane/min, after which it rises more steeply. The overall average travel time 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.

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. In order to reveal more detail in the region with most data, Figure 2 and later similar figures includes a middle panel showing data within a restricted vertical range.

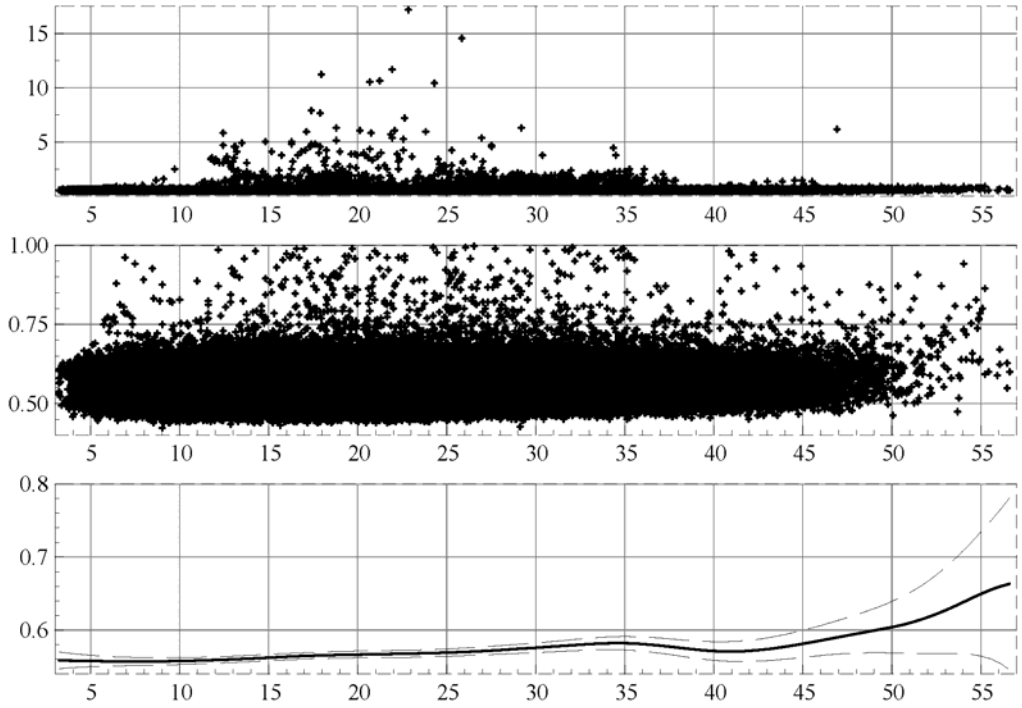

Figure 2 Travel time against hourly average entry flow

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.

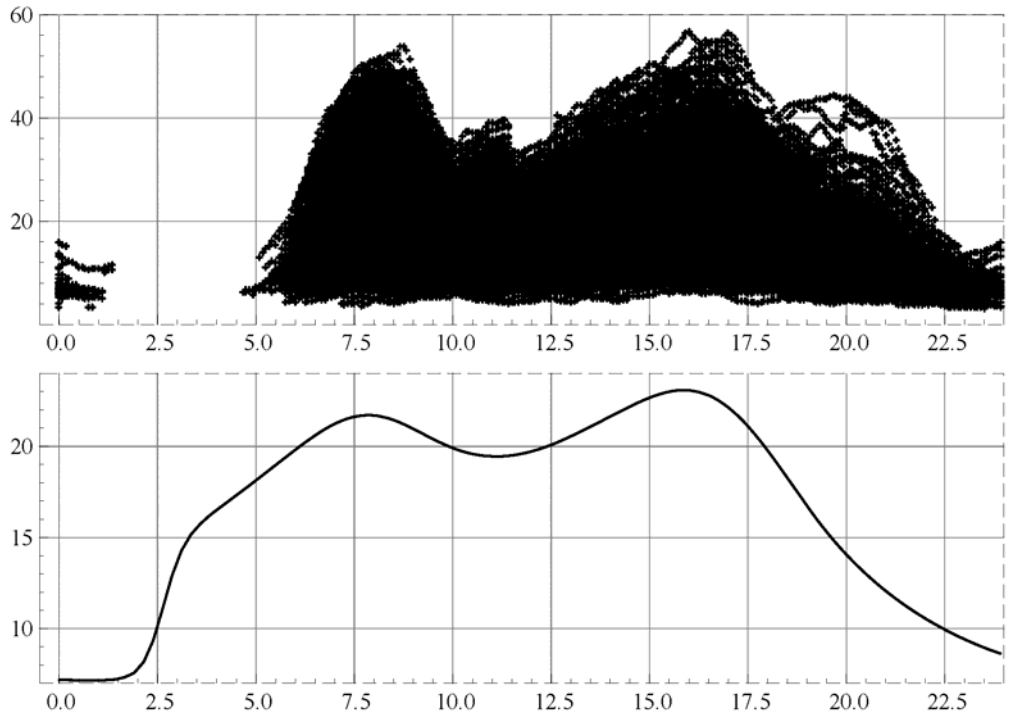

Figure 3 Hourly average flow against time of day

4. Empirical results

4.1 Model specification

We need to specify the model in (7), describing the current travel time on a link as a function of lagged travel times on the same link, entry flow via $f(E_t^n)$, queue via $h(Q_t^n)$, upstream blockage via $g(D_t^{n+1})$, and controls W and H.

We begin by discussing the controls. First, travel time on a link is affected by weather and other conditions not related to flow. We utilize the log travel time experienced concurrently on the same roadway in the opposite direction to control for these effects and label this variable W. This variable is averaged over three periods around the current time interval. Second, the speed for trucks is restricted; hence we include a variable H equal to the share of heavy vehicles measured in pce in the current exit flow. We allow for link-specific coefficients on W and a single coefficient on H.

According to the discussion in Section 2, we must regard entry flow, queue, and upstream blockage as endogenous since these variables are all affected by congestion. We therefore need to specify instrumental variables that are correlated with the endogenous variables but uncorrelated with the current residual in (7). In addition to the controls and two lags of travel time, already included as variables in (7), 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 upstream density directly, but they are unlikely to be correlated with the residual in (7) because blockages seldom 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.

We estimate link-specific fixed effects as well as link-specific parameters for *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

	lnT	lnE	Q	$\ln D^{n+1}$	H	W
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	2.842	4.226	81.41	4.588	0.9183	19.40
Minimum	-0.8519	0.8310	0.000	0.08948	0.1106	0.4208
Std. Dev.	0.1374	0.4997	10.08	0.5477	0.1428	0.3071
Skewness	5.836	-0.4590	1.991	0.01343	-0.3641	22.45
Kurtosis	78.01	2.871	7.275	2.972	2.409	723.4

4.2 Estimation results

A range of models has been estimated. We present estimates from three models based on piecewise linear specifications of functions *f*, *g* and *h*. The breakpoints for these functions were found by visual inspection of the curves resulting from estimating third order polynomials. (We prefer not to rely on the models using polynomials since our data are heavily concentrated at low

to intermediate flows, whereas we are most interested in the properties of the function at high flows; the fit of the polynomial where data are dense will determine the shape where data are sparse and hence results may be misleading.)

The function f relating to entry flow has a breakpoint at 40 pce/lane/min, which corresponds to the point in Figure 2 where travel time seems to begin to rise and just slightly exceeds the Danish design standard for lane capacity.[4] The function g for the 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 (6). 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. All 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, indicate little autocorrelation of the residuals.

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 $1/(1-0.7)\approx 3.3$ to get the values that apply when the model is solved in steady state, as shown in equation (8).

[4] 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	
Variable	Coeff.	t-Stat.	Coeff.	t-Stat.	Coeff.	t-Stat.
Const.	-0.246	-34.5	-0.234	-32.7	-0.257	-35.0
T_{-1}	0.384	83.8	0.382	83.6	0.381	83.3
T_{-2}	0.317	70.2	0.318	69.6	0.312	69.3
lnE	0.006	3.2	0.061	5.9	0.007	3.6
$(\ln E-\ln(40))*1_{\{E>40\}}$	0.267	4.8	0.226	4.5	0.194	3.8
(Q-20)*(Q>20)	-0.001	-4.1				
(Q-40)*(Q>40)	-0.001	-2.0				
$\ln E_{-1}$			-0.058	-6.1		
$\ln F_{-1}$			-0.001	-1.0		
$(\ln D^{n+1}-\ln(50))*1_{\{Dn+1>50\}}$	2.124	3.7	2.667	4.6	3.446	5.9
H	0.041	12.9	0.038	11.9	0.043	13.7
link-specific constants	yes		yes		yes	
link-specific const's * *W*	yes		yes		yes	
Number of observations	66603		67047		67470	
Adjusted R-squared	0.514		0.514		0.512	
Sum squared resid	629.927		637.689		647.261	
Durbin-Watson stat	2.124		2.112		2.094	
Second-stage SSR	625.933		629.245		631.836	

Note: $1_{\{\ \}}$ denotes the indicator function for the event in the curly brackets.

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.[5] With nine links included, there are nine link-specific effects of control variable *W*. The latter control variables almost all have statistically significant effects, typically in the range of 0.02–0.09, indicating that travel time on

[5] 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 $\ln(T_T/T_C)=\beta_H/(1-\beta_1-\beta_2)$ or $T_T/T_C = \exp[\beta_H/(1-\beta_1-\beta_2)] \approx 1.13–1.15$. This may be somewhat too small, as the speed limit for trucks is 80 km/h compared to 110 or 130 km/h for cars. Actual speeds tend to be higher. 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.

the opposing link affects travel time on the link in question with an elasticity of roughly 0.02–0.09.[6]

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, which is not altogether surprising given the discussion in section 2.1. In particular, the entry flow is not measured but is approximated from 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 increasing entry flow having a significant positive effect on travel time. The coefficients imply an elasticity of travel time with respect to entry flow of about $0.008*3.3 \approx 0.026$. For entry flows larger than 40 pce/lane/min the coefficient is similarly significant, it is larger as expected, implying an elasticity of travel time with respect to entry flow of about $0.254*3.3 \approx 0.83$. The coefficient for the downstream density implies a long-run elasticity of travel time with respect to downstream density of $3.446*3.3 \approx 11$, which seems to be a substantial effect. Our results thus confirm that queue spillbacks can be an important contribution to congestion, as has long been assumed throughout the engineering literature.[7]

[6] 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.

[7] See, for example, the freeway simulation models described by May (1990).

There is, however, an empirical weakness in our results that may indicate the specification is not fully satisfactory. Despite our having about 67,000 observations, the three key coefficients for congestion determination — those on the two variables using entry flow and the one using downstream density — are only moderately "significant", with asymptotic t-statistics between three and six. Furthermore, we recognize that the specification with two lagged values of travel time is only one of many types of dynamics that could be present. If we re-estimate the same model but allowing for first-order autocorrelation, we find it difficult to achieve convergence; but it appears that these key coefficients are not stable and the estimated autocorrelation, although small (~-0.1), is statistically significant. Thus, the results that matter most to our subsequent calculations of external costs are sensitive to the assumed time-series properties of the model.

5. Calculation of marginal external costs

This section presents the methodology and the results of the calculation of marginal external cost (*mec*). We begin by observing that with entry flow E and travel time T, the internal cost (ignoring monetary costs) is T and the total cost of all users is TE. Then we may find the *mec* by differentiating total cost with respect to entry flow and subtracting the internal cost. This yields

$$mec = \frac{\partial (TE)}{\partial E} - T = \frac{\partial T}{\partial E} E. \tag{4.1}$$

We now turn to our preferred empirical model. Ignoring constants and the error term, and solving for the steady-state travel time, we can write it as follows, where a variable lacking a time subscript is a steady-state value:

$$\log T^n = \gamma_1 \log E^n + \gamma_2 \log\left(\frac{E^n}{40}\right) 1_{\{E^n > 40\}} + \gamma_3 \log\left(\frac{D^{n+1}}{50}\right) 1_{\{D^{n+1} > 50\}} \tag{4.2}$$

where $1_{\{\ \}}$ denotes the indicator function for the event in the curly brackets.

We note from (4.2) that the dependence of travel time on flow in (4.1) works through two paths: first through flow on the current section (terms involving γ_1 and γ_2), and second through the downstream density (term involving γ_3). Thus *mec* in (4.1) is the sum of two

components, mec_E and mec_D. We calculate both components by differentiating the relevant terms of (4.2) with respect to E_n. The first term is straightforward:

$$mec_E^n = \left(\gamma_1 + \gamma_2 1_{\{E_n>40\}}\right)T^n . \tag{4.3}$$

The second term requires the derivative

$$\frac{\partial D^{n+1}}{\partial E^n} = \frac{\partial (T^{n+1}E^{n+1})}{\partial E^{n+1}}\frac{\partial E^{n+1}}{\partial E^n} = \left(\frac{\partial T^{n+1}}{\partial E^{n+1}}E^{n+1} + T^{n+1}\right)\frac{\partial E^{n+1}}{\partial E^n} . \tag{4.4}$$

For simplicity we ignore the time lag and the compression or expansion of traffic due to changing travel times, and simply assume that conditions are steady over the length of time required for the surge in E^n to become a surge in E^{n+1}; hence $\partial E^{n+1}/\partial E^n = 1$. Therefore the portion of equation (4.1) coming from the term involving γ_3 is:

$$\begin{aligned} mec_D^n &= \gamma_3 \frac{T^n E^n}{T^{n+1}E^{n+1}}\left(\frac{\partial T^{n+1}}{\partial E^{n+1}}E^{n+1} + T^{n+1}\right)1_{\{D_{n+1}>50\}} \\ &= \gamma_3 \frac{D^n}{D^{n+1}}\left(mec_E^{n+1} + T^{n+1}\right)1_{\{D_{n+1}>50\}} \end{aligned} \tag{4.5}$$

Consistent with the assumption that conditions are changing only slowly, we assume $D_n / D_{n+1} = 1$. We then can compute the *mec* for each observation in the sample as simply

$$\begin{aligned} mec &= mec_E + mec_D \\ &= \left(\gamma_1 + \gamma_2 1_{\{E>40\}}\right)T + \gamma_3\left(mec_E + T\right)1_{\{D>50\}} \end{aligned} . \tag{4.6}$$

We compute both components of (4.6) 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 (4.6) in this way is that the formula uses realisations of T including error terms. This matters for the result as the dependency of *T* on error terms in (7) 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. Note that (4.6) is discontinuous as a function of *E*, for given *T*; but since *T* is random and its expectation depends on *E*, the expectation of *mec* is a continuous function of *E*.

Figure 4 presents scatter plots and a smoothed mean of the *mec* against the entry flow, where the entry flow is 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 (more importantly) to the random term in travel time, which appears in equation (4.6) for *mec*. The discontinuity of the derivative of the fitted model is visible as a vertical gap between a cloud of low *mec*'s and a cloud of larger *mec*'s. The smooth of the *mec* may be considered to be an estimate of the expected *mec* conditional on the hourly average entry flow.

The *mec* is initially small and rises slowly until an entry 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. From this point, the *mec* rises more steeply and at a flow of 40, *mec* has reached 0.3 min/km or about a little more than half the average travel time. At 50 pce/lane/min, *mec* has risen to about 0.7 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.

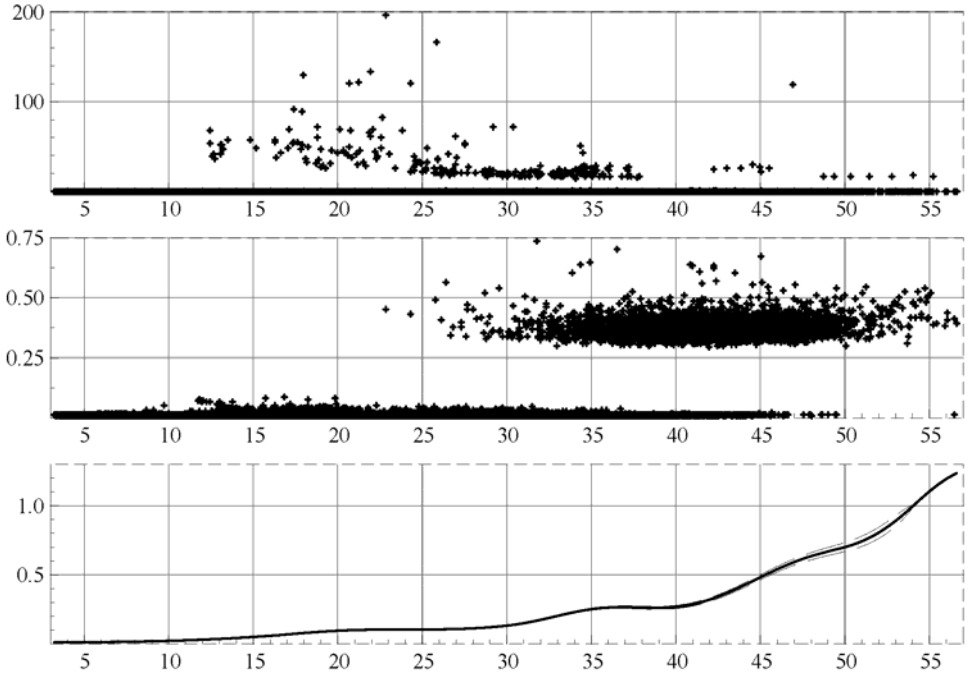

Figure 4 mec against hourly average entry flow

Figures 5 and 6 present the two components of the *mec*. It seems the component reflecting downstream congestion, mec_D, is extremely variable and its average even dominates at flows up to about 40 pce/lane/min, after which the component reflecting current congestion, mec_E, is generally the larger.

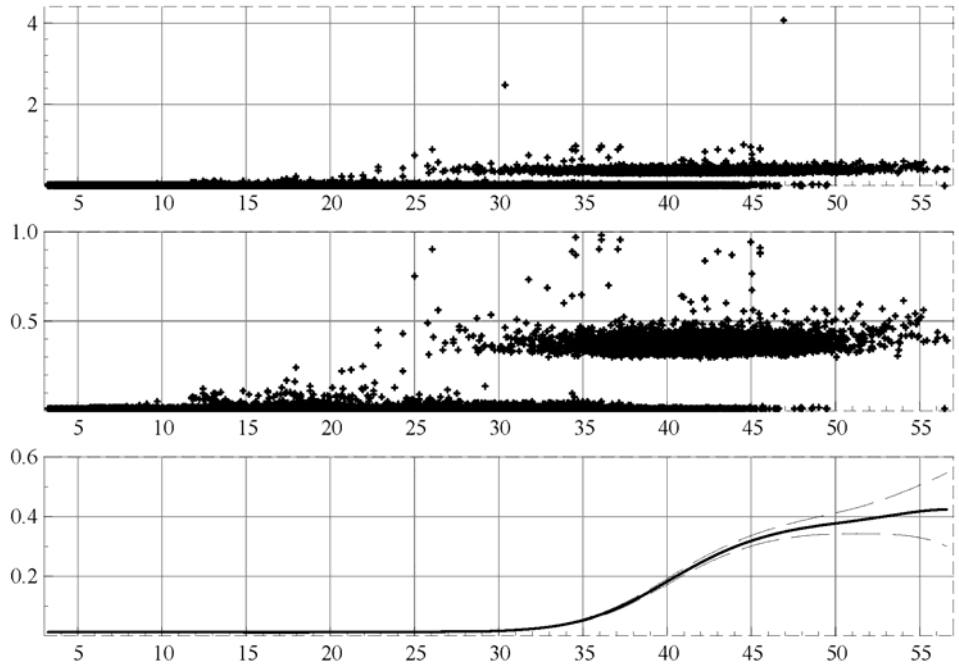

Figure 5 mec_E against hourly average entry flow

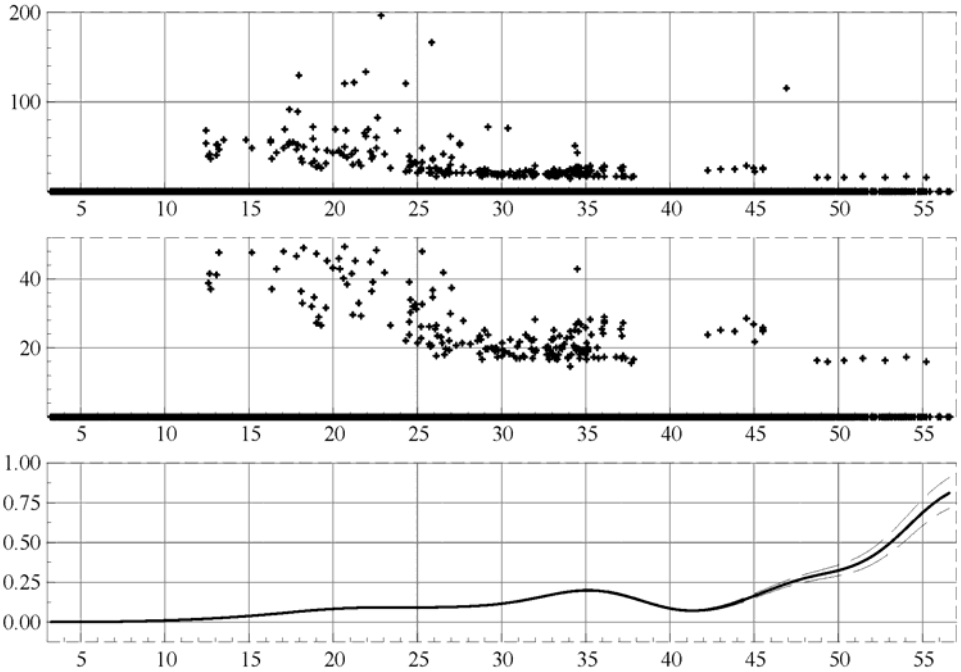

Figure 6 mec_D against hourly average entry flow

Finally, Figure 7 shows the *mec* against the time of day. Its average follows the peaks in traffic and seems to be highest at about 0.2 min/km at around 3 p.m. Evidently, on this network the lower-flow situations are most common 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.

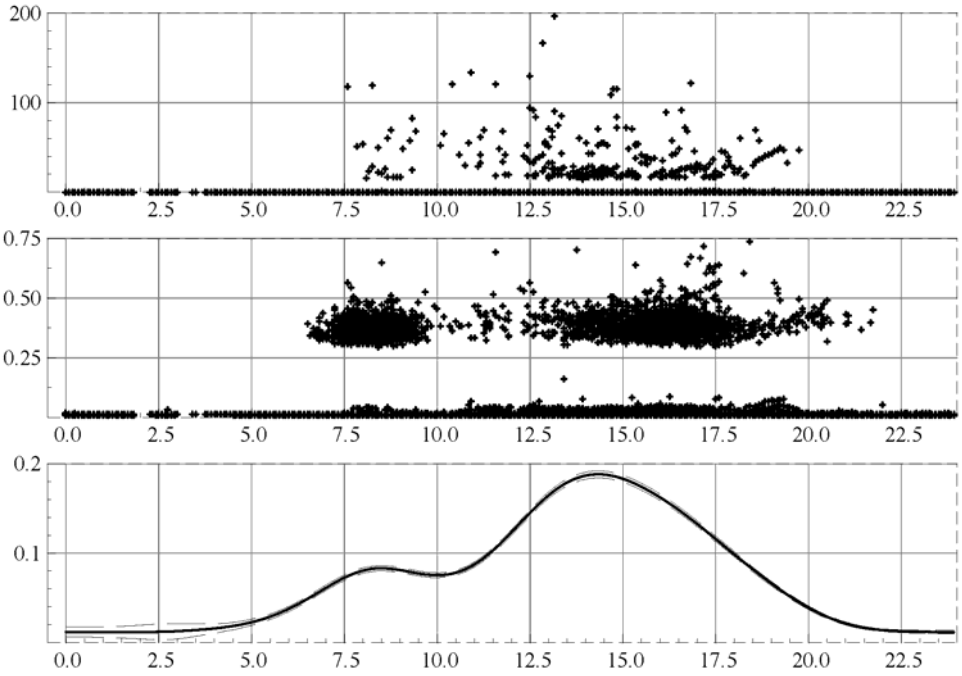

Figure 7 *mec* against time of day

We noted in Section 4 that our empirical results are sensitive to the specification of autocorrelation in the error terms. Furthermore, while the lagged dependent variables in (7) are intended to capture dynamic effects involving internal queues operating in a regime of hypercongestion, they could also be proxies for autocorrelation resulting from persistent influences not captured by our control variables *W* and *H*. In that case, the time-series correlations in the data would not be explained by congestion dynamics, and thus would not raise the external costs. The effect of these two factors is to greatly reduce the marginal external costs. Our best assessment is that lagged travel times have a robust effect and therefore the dynamics are real, but that the magnitude of the direct effect of entry flow and downstream density are uncertain due to specification uncertainties.

6. Conclusions

This paper has contributed to the measurement of the marginal cost of freeway congestion in several ways. Simultaneity of speed and traffic flow is likely to be important when there is congestion. An indication of this issue is found in Figure 2, which seems to show that average travel time may even decrease as flow increases. This would imply a negative marginal external cost of adding vehicles to flow which is blatantly nonsensical.

Using our structural model, we have argued for the use of observations from other links in the same network as instruments for flow in an equation describing travel time as a function of flow. Thus we believe we are able to go some way in tackling the simultaneity issue. Our results using these instruments indicate a positive and increasing marginal external cost in accordance with the a priori expectation. But we acknowledge the sensitivity of these results with respect to specification of the time-series properties of the residuals.

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.

The resulting model seems plausible. So do the estimates of the marginal external cost at various levels of flow and during the day. From the structural model we derive an expression for the marginal external cost that is easy to compute for each observation in the sample. The results of this computation indicate large dispersion of the marginal external cost, induced by variation in the observed travel times. The positive skewness of the distribution of marginal external cost reflects the positive skew of the distribution of travel times.

The distribution of travel times and flows is very important. There are strong dynamics that are difficult to disentangle. We suspect that our instruments are inadequately accounting for the endogeneity of entry flow, causing our model to poorly explain the many data points observed in Figure 2 that have very high travel times yet only moderate flows. These observations probably result from internal queues which block entry, as suggested by our structural model, but we have not found a satisfactory way to measure them and include them in an empirical reduced-form specification.

What seems clear to us is that a satisfactory empirical model of congestion needs to consider both internal dynamics on a given link and feedbacks to and from adjacent links. Only

then can one really speak precisely of the marginal cost of adding cars to a network at a given place and time.

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

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.

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.

Notation

F = exit flow (pce/min per lane)
E = entry flow (pce/min per lane)
T = travel time (minutes/km)
D = density (pce/lane-km)
q = size of internal queue on link (pce/lane)
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)
Δt = width of time interval, min (=5 in our data)
t = time period (integer)
n = section (larger numbers are downstream)

Appendix: Approximating entering flow from observed upstream flows

The relevant upstream flows are those during the current and immediately previous time periods, in the case of short sections (those that take less than five minutes to traverse); and those during once and twice lagged time periods, in the case of longer sections. (No section takes longer than two time periods to traverse, so we need not consider three lags.) Thus we construct a flow variable equal to a weighted average of those three observed flow, with the weights equal to the proportions of vehicles that could be expected to have been observed during the current and previous time period, respectively:

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

In order not to introduce endogeneity into the flow variable, we compute these weights using the average speed on the entire network, S*, expressed in km/min. Consider the vehicles exiting link *n* during the five-minute time interval *t*, which we take to begin at time 0 and end at 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 $w_t=1-w_{t-1}$. 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 $w_{t-1}=1-w_{t-2}$. We can summarize for both cases as follows:

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