**Published paper**

A. D. May, S. P. Shepherd, and J. J. Bates (2000) Supply Curves for Urban Road Networks. Journal of Transport Economics and Policy, 34(3) 261-290

Journal of Transport Economics and Policy, Volume 34, Part 3, September 2000, pp.261-290

# Supply Curves for Urban Road Networks

**A. D. May, S. P. Shepherd, and J. J. Bates**

Corresponding author: Professor Tony May, Institute for Transport Studies, University of Leeds, Leeds LS2 9JT UK. S. P. Shepherd is also at ITS Leeds. J. J. Bates heads John Bates Services at Oxford.

This study was supported by a grant from the Engineering and Physical Sciences Research Council, whose support is gratefully acknowledged. The proposal for the study arose from a series of discussions with colleagues in consultancy and universities, and the research has benefited considerably from advice from a number of colleagues. The paper has also benefited from the comments of referees. The authors are grateful for their contributions, but stress that the conclusions are the authors' own, and do not necessarily reflect the views of others.

**Abstract**

Supply curves are essential for demand prediction in aggregate networks, and need to relate cost of use to flow demanded rather than to flow performed. The paper defines the parameters of relevance and the relationships required. A microsimulation model is used to generate such relationships for hypothetical networks. Network performance curves can exhibit backward-bending relationships between speed and flow performed, but these curves are inappropriatefor demand prediction. Relationships between speed and flow demanded are monotonic, but not necessarily linear. They can exhibit spatial and temporal dependencies, both of which requires them to be re-estimated if demand patterns change.

*Date of receipt of final manuscript: May 1999*

# Introduction

Fundamentally, the prediction of the level of use of an urban network, and of the costs of using it, depends on an understanding of two relationships. The first is the demand curve, which determines the demand for use of the network as a function of the cost of using it, which may be measured solely in terms of time, or of some combination of time and operating costs. The second is the supply curve, which determines the cost of using the network to individual vehicles (again measured in terms of time, or time and operating costs) as a function of the total demand. The intersection of these two curves determines the level of use of the network that will occur in practice, and the resulting costs to the vehicles that use it.

For an individual link, demand can be measured in terms of flow — the number of vehicles entering the link in a defined period. The supply curve is often based on the inverse of a speed-flow curve, thus limiting the costs considered to travel time per unit of distance. Speed-flow relationships for individual links and junctions are well accepted and are used extensively to evaluate the benefits of investments in additional capacity (see, for example, DoT, 1996). However, the use of link-based relationships to analyse demand and supply in complex urban networks is time-consuming. It is also potentially inaccurate, since such relationships may ignore the interactions between links at a junction.

An alternative approach, which has been adopted in several studies of road pricing (Thomson, 1967; Harrison *et al.,* 1986; Evans, 1992) and in recent developments in strategic transport modelling (Bates *et al.*, 1991; Oldfield, 1993) is the use of more aggregate supply functions, and specifically area speed-flow relationships. In moving from a link to an area, the simple definition of "flow" as the number of vehicles per unit of time is no longer applicable, since the time spent or distance travelled by a vehicle within the area is clearly relevant. For this reason, aggregate traffic "flow" is typically measured in veh-km/h, and the corresponding speed calculated as (veh-km)/(veh-h). Such aggregate supply models can be simple "single-link" models in which speed and flow are related on a link that represents the performance of a whole area, or they can relate separately to different zones in a large urban area.

The general form of such relationships is frequently based on work by Duncan *et al.* (1980), who used data from the UK urban congestion surveys of 1963 to 1976. Their relationship was linear between speed and the increase in flow above off-peak conditions. It also included a term in junction frequency for central areas, and one in density of development for non-central areas. These results were incorporated into TRL's London Area Model (LAM) (Oldfield, 1993) as a linear relationship between speed and the change in veh-km/h. This model was used for the transport strategy analysis conducted for the London

Planning Advisory Committee (May and Gardner, 1990). However, the relationship had to be modified by introducing an arbitrary increase in slope below a specified speed; without this, the model predicted large increases in traffic in central and inner London with only small speed reductions, which was not credible.

A relationship similar to this has since been employed in the START model (Bates *et al.*, 1991), which has been used for integrated transport studies in a number of cities (May and Roberts, 1995). In the START applications, each zone in the study area is allocated one or more area speed-flow relationships that may, for example, be defined separately for inbound, outbound, and orbital travel. Because the model is a marginal one, the relationship is between change in speed and change in flow, measured in (veh-km/veh-h) and (veh-km/h) respectively. These relationships can be determined from empirical data or from more detailed network models such as SATURN (van Vliet, 1982).

The performance of policies is likely to be particularly sensitive to the shape of these relationships at and around the lowest speeds that are observable in practice. Such speeds reflect the equilibrium between supply and demand in congested conditions, and mis-specification of the supply relationship in these conditions will lead to over- or under-estimation of the benefits of congestion relief.

However, this makes it difficult to develop reliable relationships from empirical data. In practice, average network speeds less than around 15 km/h are rarely observed, because congestion at this level becomes self-regulating. Yet the lack of data below this speed makes the shape of the relationship in this area particularly uncertain. Relationships can be generated from network models, which do predict lower speeds, but the lack of such conditions in reality calls into question the assumptions in the network models. More fundamentally, as discussed in the next section, it is questionable whether it is appropriate to employ such relationships; what is needed is a relationship between speed and the flow demanded rather than between speed and the flow that occurs in practice.

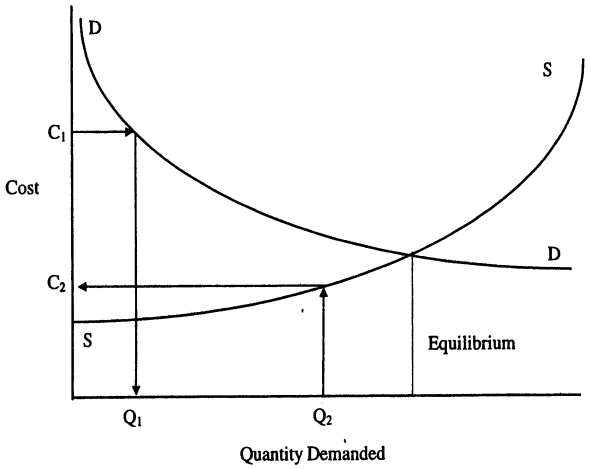
This paper presents the results of a study designed to investigate this problem, and to develop more defensible relationships using microsimulation. The study, which was funded by the UK Engineering and Physical Sciences Research Council, was prompted by experience with the use of the START model, and commenced with a series of discussions to specify more precisely the parameters and relationships required. These are defined and justified in the next section. The third section describes the modelling approach, which used the microsimulation model NEMIS (Mauro, 1991). The fourth section presents the results; the concluding section discusses their implications for strategic modelling and for policy.

# Theoretical Considerations of Demand and Supply

**Demand and supply on a single link**

We consider first the case of a single link. Here it is possible to appeal to reasonably firm definitions of capacity, in terms of the link's ability to discharge traffic. Provided capacity is not exceeded, the traffic will get through, though there may be a tendency for average speed to fall. The time taken to traverse the link may increase somewhat and there will, except in steady-state conditions, be some differences between inflow and outflow, but these will have no significant effect on the throughput. Effectively, the flow demanded is equal to the flow performed until capacity is reached.

**Figure 1**
*Standard Model of Supply-Demand Equilibrium, Showing Directions of Causality*




When $q$, the flow demanded, exceeds capacity $K$, queues will form, and the time taken to exit from the link will depend on the length of the queue. In practice, we do not experience infinite queues, because the periods over which $q > K$ are strictly limited: once $q$ falls below $K$, the "server" regains the capacity to reduce the queue, eventually to zero. It is therefore essential, when discussing the effects of demand in excess of capacity, to define the period over which the demand applies.

From the point of view of an equilibrium between supply and demand, we are concerned with how the demand will change as the "cost" of the journey increases, and to answer the question: given that so many people wish to make the journey, what will the conditions be? There are two important observations to be made in connection with the conventional supply and demand diagram (Figure 1). Whereas the X-axis is typically described as the quantity axis, it relates rather to the quantity demanded. The two curves, conveniently represented on a single graph, actually have very different implications. The demand curve, contrary to standard mathematical convention, gives the value along the X-axis (demand) consequent on the Y-value (cost); thus cost $C_1$ would give rise to a demand $Q_1$. The supply curve works in the conventional direction: given a demand of X, at what cost can this be met; a demand $Q_2$, if satisfied, would result in cost $C_2$. The implication is that the supply curve needs to be in units of the quantity demanded.

**The "backward-bending supply curve"**

We now go on to discuss a major source of misunderstanding, which appears to go back at least to the seminal article by Walters (1961).

For demand below capacity, there is abundant empirical evidence supporting some kind of speed-flow relationship. There is less certainty about what happens when capacity is exceeded (see Hall and Montgomery, 1993). The traditional approach has been to propose a relationship that postulates that the speed of traffic through the link is a declining function of the density (in veh/km) of traffic on the link. The flow of traffic through the link (that is, the rate of discharge per hour from the link) is then given by the product of the speed and the density. Such a functional form has the general properties that (a) there is a maximum flow, (b) there are in general two possible values of speed associated with a given flow, (c) flow is zero at zero speed, and (d) speed is zero at maximum density (Greenshields, 1934). Figure 2 illustrates the well-known parabolic form of this relationship.

There is a good deal of controversy about the extent to which such a curve represents an empirical phenomenon, as opposed to being merely a convenient idealisation. Hall and Montgomery (1993) have recently argued that the curve is really a composite of three separate conditions and they note that "it is unlikely that the whole of [the curve] can be observed in any one location." The empirical evidence for the low speed portion of the curve must be judged as relatively weak, but we may certainly accept it as a possibility.

It is also important to point out, as Small (1992) notes, that the underlying theory relating speed and density is essentially about instantaneous effects. In practice, of course, the ideas are applied to average conditions over a defined period, so that there is a temporal aggregation that abstracts from the actual profile of demand during the period.

**Figure 2**
*Parabolic Speed-flow Relationship*

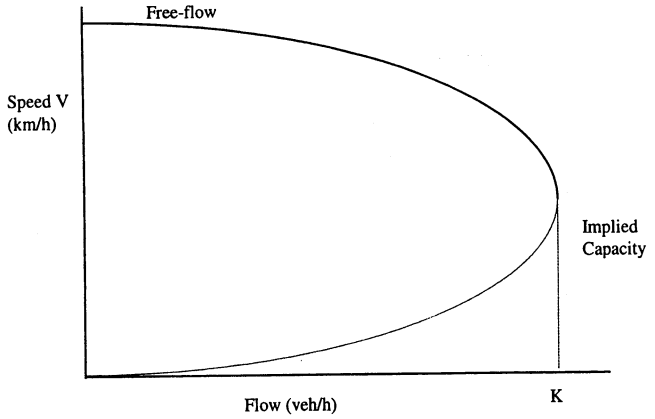
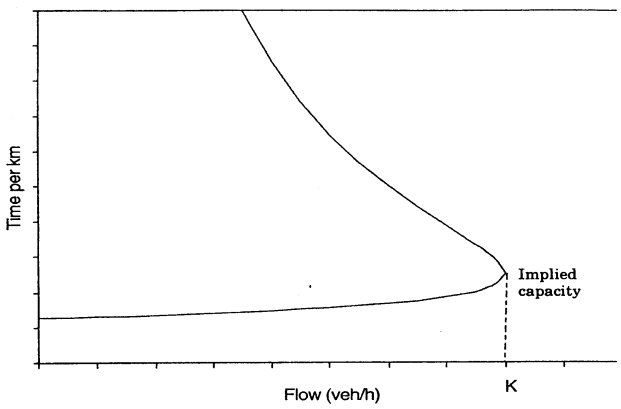



**Figure 3**
*Inverse of Figure 2*



If we re-cast the Y-axis as the inverse of speed (time per unit distance), we obtain the "backward-bending" curve shown in Figure 3. It should be clear, however, that this has nothing to do with the supply curve. This is because the quantities represented along the X-axis in Figures 2 and 3 are not in demand units. Rather, they are the expected (or observed) result of plotting the combinations of speed and actual flow that occur for different levels of demand. In other words, the X-axis is in terms of flow performed, not flow demanded. This observation has been made before (see Neuberger, 1971) but generally ignored.

Unfortunately, a significant number of economists from Walters onwards have attached particular importance to this backward-bending property, interpreting flow performed as the relevant quantity variable in supply-demand equilibrium, and labelling the equivalent of the lower portion of the speed-flow curve as the case of "hyper-congestion". For a recent discussion of this, see Small and Chu (1997).

If it is accepted that the X-axis needs to be a measure of flow demanded in a given period of time, then the Y-axis must describe the costs associated with that demanded flow. Herein lies a second problem. Conventionally data of the kind presented in Figure 2 are measured over a given period of time; they provide an engineering measure of the flow performed in that time period, the time spent in that period, and the resulting speed. However, what is needed for an analysis of the costs associated with a given demand is the time spent by all the demand flow up to the time that it exits the link. This requires that vehicles be "tracked" until they leave the link. In turn this means that the period over which this travel time is measured may well be longer than that over which the demand occurs, and over which performance associated with that demand would conventionally be measured. To avoid confusion in what follows, we draw a clear distinction between:

(1) flow performed and time performed in a defined period, which between them provide an engineering description of the performance of a link, and hence its performed speed (the inverse of time); and
(2) flow demanded in a defined period, and flow and time supplied once that demanded flow has left the link, where the latter may be measured over a longer period of time, time supplied includes time spent queuing (perhaps to enter the link), and the inverse of time supplied is referred to as a "pseudo-speed" to reflect the fact that it can include the effects of queuing.

In a single link, these distinctions are not always necessary. When the link is uncongested, flow performed and flow supplied will be identical, because they are measured in the same time period. More generally, flow supplied will equal flow demanded (unless there is some long-term constriction that stops the flow demanded being satisfied), but the period over which flow is supplied will be longer than that over which it is demanded. Similarly, time supplied may

equal time performed when there is no congestion; once there is it will differ, and be measured over a longer period. The same applies to performed speed and supplied pseudo-speed. In a network, as we shall see, these distinctions become more important.

We can summarise the position in Figure 4, where three separate quantities are plotted against demand, here assumed to be appropriately defined and measured as flow demanded in a given time period. The top part of the diagram shows the performed flow, and the flow supplied by the system. Until demand reaches capacity, both are more or less equal to demanded flow. Beyond capacity, performed flow cannot exceed capacity. It is possible that throughput, and hence performed flow, may fall slightly due to friction at the point of limited capacity. Supplied flow continues to rise with demanded flow, but as noted above, will be satisfied over a longer period. Only if there is a long-term constriction will it fall below demanded flow.

The middle part of the diagram presents the supplied time per km. This will start at the inverse of free flow speed, but will rise as capacity is reached, reflecting the interaction between vehicles on the link. Beyond capacity it will rise with demanded flow, reflecting the additional time spent queuing to enter the link. This has the form of the required supply curve, though it should be noted that strictly speaking other elements of generalised cost are required, including variations in vehicle operating cost with speed and flow conditions, as well as a possible allowance for lateness and variability.

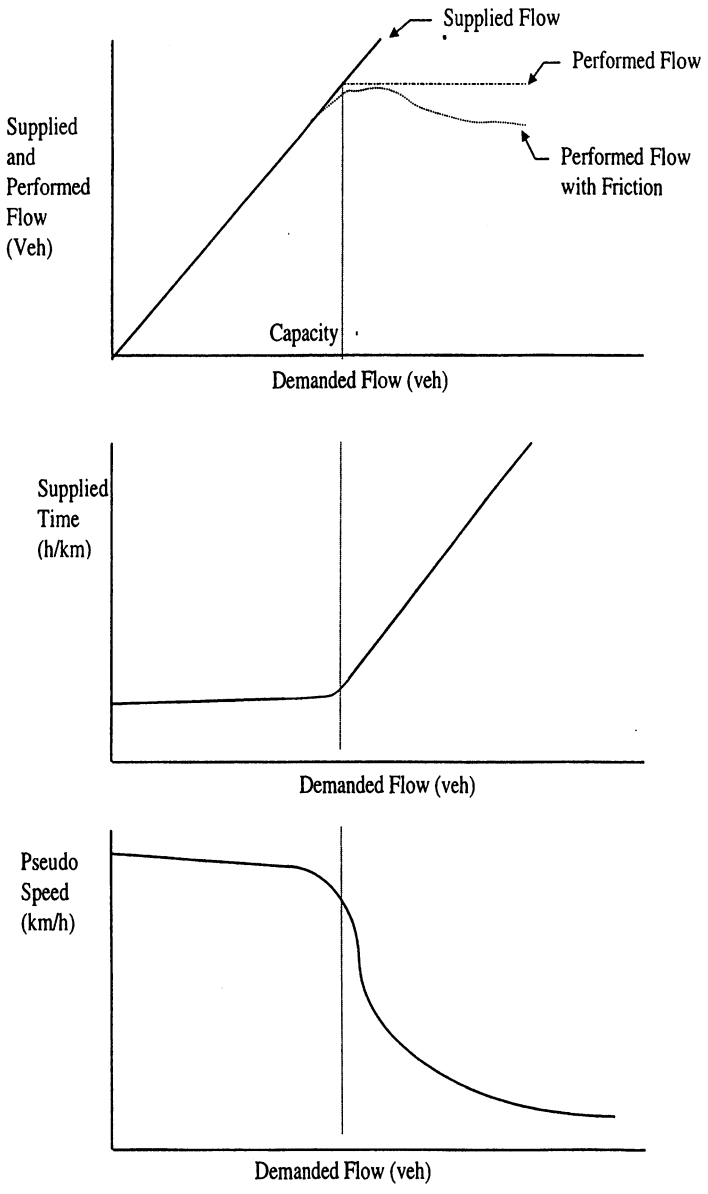
The bottom part of the diagram converts the supplied time per km to its inverse, which reflects the "pseudo-speed". This starts by following the performance speed-flow relationship, but continues to fall with increasing demanded flow reflecting the effects of queuing.

**Demand, performance, and supply for area-wide analysis**

As long as we remain with the simple case of a single link, both demand and performed flow can be measured in vehicles per hour. However, as already noted, once we move to a wider area, with a number of links, and different origin-destination demands, we are forced to look for an aggregate quantity that is appropriate for the totality of movement in the specified area. The common approach of extending the definition of travel demand to the total vehicle-km per hour demanded in the system is a useful abstraction, even if it is at some remove from the actual factors affecting the demand for travel.

There are, nonetheless, some important caveats, as highlighted by Hills (1993). In the first place, different combinations of origin-destination demands could yield the same demands in vehicle-km terms, while resulting in different levels of vehicle-hours spent, so that we cannot expect the system to be generally independent of the actual pattern of O-D movements.

**Figure 4**
*Key Relationships with Demand*
*Single Link Case*



Second, we would not wish to define the demand units in a way that is sensitive to the actual routeing chosen. It seems reasonable to assume that there is an ideal routeing pattern at low levels of demand (thus based on free-flow speeds and paths), and that the associated distances will serve for the calculation of demand in veh-km per hour. The fact that in practice different routes (typically, longer) are chosen as the system becomes loaded has nothing to do with the definition of demand, but is once again connected with system performance. It would, alternatively, be possible to use crow-fly vehicle kilometres as the measure of demand. It has been suggested that it would be preferable to base demand on trips, since people do not desire distance *per se*. However this, too, would need to be associated with a particular matrix, since a matrix with longer trips would place greater demands on the network. We therefore define demanded flow or, more strictly, demanded travel, as the vehicle-km per hour, for a given matrix distribution, with each origin-destination pair assigned to the free flow paths.

Performed travel is less controversially defined as the vehicle-km per hour observed on the network in a defined period.

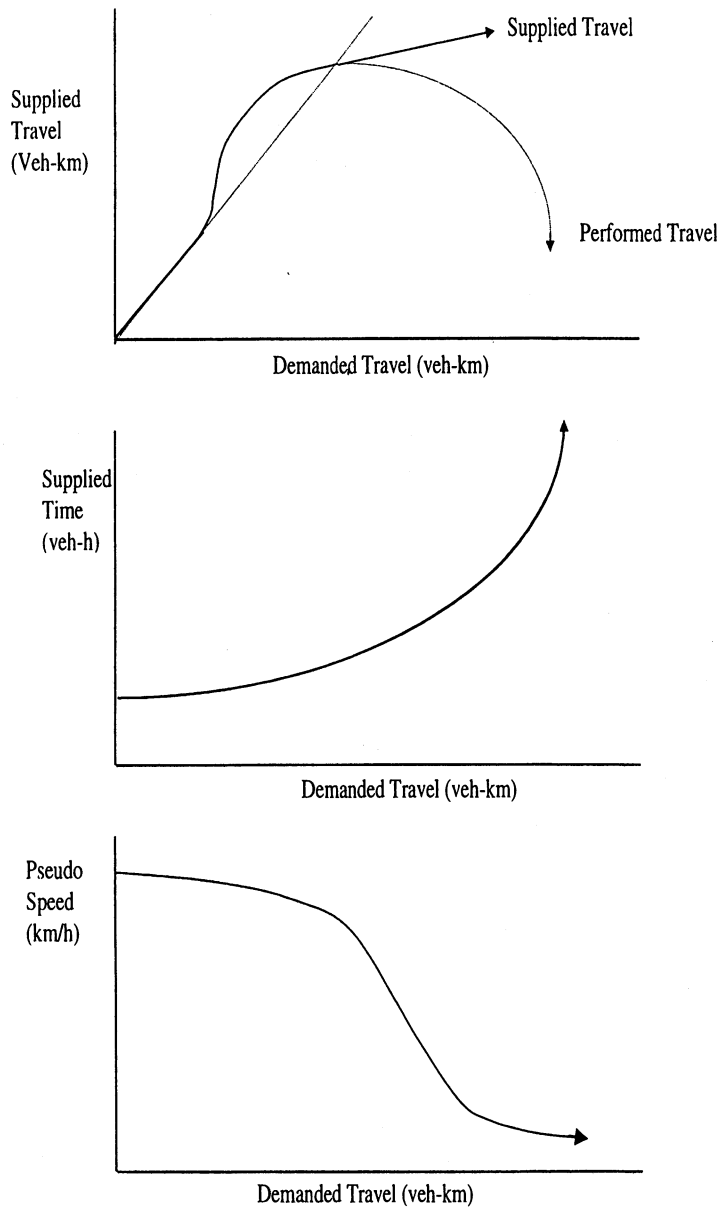
Supplied travel needs to be determined by tracking all the trips demanded in a given period until they have reached their destinations. It should be noted that, at any level of demand, this tracking process does not end at the same time for all origin-destination pairs; some will take longer to complete their journey and, in the extreme where the network becomes grid-locked, some may never complete their journeys.

Supplied time is then the sum of the time taken by all of the vehicles in the matrix in completing their journeys. In the extreme, where grid-lock occurs, this can tend to infinity. Supplied time per km is then obtained by dividing this by the demanded travel. The supplied pseudo-speed is then the inverse of this.

We now proceed to construct the area-wide counterpart of Figure 4, which we show schematically in Figure 5. In the upper part we show the performed and supplied travel against the demanded travel. Well before demand reaches "capacity", we can expect performed and supplied travel to rise faster than demanded travel, because of re-routeing. It should be noted that "capacity" becomes altogether more difficult to define, and is likely to be dependent on the O-D pattern and the connectivity of the network. However, it can be expected that performed travel will reach a maximum. Beyond this it may fall dramatically, if part of the network becomes blocked, or it may be maintained at a slightly suboptimal level if, despite queuing to enter the network, traffic can continue reaching all destinations.

Supplied travel will continue to increase with demanded travel, but may conceivably fall later if the additional demand leads to the network becoming blocked.

**Figure 5**
*Key Relationships with Demand*
*Network Case for a Given O-D Pattern*



In the middle part of Figure 5 we show the supplied time per km against demanded travel. Here supplied time per km is measured by determining the total time spent by all vehicles in the matrix until they reach their destination, and dividing it by the demanded travel. This will start as the inverse of the free flow network speed (for the given O-D pattern) and then rise as friction between vehicles and traffic streams increases. However, re-routeing opportunities will result in it rising less rapidly than it would were routes fixed. What happens beyond this will depend on whether part or all of the network becomes jammed. In the extreme supplied time per km could approach infinity.

The bottom part of Figure 5 shows the supplied pseudo-speed, again as the inverse of supplied time per km. This will start at the free flow network speed, follow the performance speed-flow relationship initially, and then continue to decline, with increasing demanded travel. In the extreme it could approach zero if the network becomes jammed.

**Parameters and notation**

The crucial points from these theoretical considerations can be summarised as follows:

- a distinction needs to be made between engineering measures of the performance of a network in a given period, and supply measures, which need to be assessed over the period during which demand is satisfied;
- the "backward bending supply curve" is in practice a performance curve and has no role to play in an economic analysis;
- for area-based analysis, veh-km/h is an appropriate unit for measuring the travel performed and supplied;
- veh-km/h is a somewhat less appropriate unit for measuring travel demanded in an area, and needs to be defined on the basis of a given set of (free-flow) paths;
- veh-km/h and veh-h/h supplied should be measured over the period during which trips occur, rather than over pre-defined time-slices.

This analysis suggests the use of the following measures for the analysis of networks:

- performed travel ($VK_p$, measured in veh-km/h): the total veh-km performed in the network over a defined period, divided by the length of that period in hours;
- performed time ($VH_p$, measured in veh-h/h) : the total veh-h performed in the network over a defined period, divided by the length of that period in hours;
- performed speed ($U_p$, measured in km/h) : the ratio $VK_p/VH_p$;
- demanded travel ($VK_d$, measured in veh-km/h) : the total veh-km with distance measured along free-flow paths, relating to vehicles wishing to enter

the network over a defined period, divided by the length of that period in hours;

- supplied travel ($VK_s$, measured in veh-km/h) : the total veh-km, with distance measured along actual paths, travelled in the network by vehicles entering, or queuing to enter the network in a defined period, divided by the length of the appropriate period in hours;
- supplied time ($VH_s$, measured in veh-h/h) : the total veh-h spent in the network by vehicles entering, or queuing to enter the network over a defined period (including time spent queuing on entry), divided by the length of the appropriate period in hours;
- supplied time per km ($T_s$, measured in h/km): the ratio $VH_s/VK_d$.

## The Modelling and Data Collection Methodology

The measures described above may be constructed by collecting the vehicle-km and vehicle-hours for each of a set of demand levels ranging from relatively free-flow to a fully congested regime using a suitable traffic model. This study employed a micro-simulation model, NEMIS (Mauro, 1991), to ensure that the effects of increased congestion on the interactions between links and between junctions were fully represented. NEMIS simulates individual vehicles, which interact via car following rules, lane changing and gap acceptance rules, and can be used to simulate signalised and non-signalised junctions.

**Definitions**

Within each "generating period", NEMIS generates vehicles according to a set of trip matrix demands $D_{ij}$ for all $ij$ pairs. A basic trip matrix is defined that is small enough to maintain free-flow conditions. To represent different levels of demand, the basic trip matrix is factored up through a number of simulations by a quantity $F$ (typically 10-20 such simulations). For each level of $F$, vehicle-kilometres and vehicle-hours are collected in the micro-simulation model NEMIS for individual vehicles and aggregated over zones, link types, and time periods.

The required quantities need to be calculated using two distinct measurement approaches. The "time-slice" approach divides the simulation into equal time periods to give the vehicle-km and vehicle-hours performed in a period. The tracking approach is used for creating supply measures: it tracks each vehicle through the network, including time spent queuing on external links at the entry points to the network. The data is then aggregated by generating period,

but measured over the period during which the trips generated take place, up to a limit imposed by the end of the simulation period.

**The time-slice approach**

The time-slice approach is used to obtain performance measures for an area or link type in a network over a specified period (for example, an hour), referred to as the given performance period. Let us define the following:

$\pi$ the performance period;
$H_\pi$ the length of the performance period;
$t_\pi$ the start of the performance period $\pi$, which will end at $t_{\pi+1} = t_\pi + H_\pi$;
$\lambda$ the basic simulation time increment (1 second in NEMIS);
$L_{sk}$ a set of links of type $s$ (inbound, outbound or orbital) within zone $k$;
$V_l(t)$ the list of vehicles at time $t$ on NEMIS link $l$;
$n$ an index for a vehicle;
$d_n(t)$ distance travelled by vehicle $n$ from time $t$–$\lambda$ to time $t$;
$h_n(t)$ $\lambda$ (= 1 second) if vehicle $n$ is present at time $t$, otherwise zero.

With these basic definitions we can now define the rate at which vehicle-km are performed in period $\pi$ on links of type $s$ within zone $k$ as follows:

$$VK_p^{\pi sk} = \left( \sum_{t=t_\pi}^{t_{\pi+1}} \sum_{\forall l \varepsilon L_{sk}} \sum_{\forall n \varepsilon V_l(t)} d_n(t) \right) / H_\pi \quad . \tag{1}$$

We have summed the distance travelled in each time increment by vehicles present on links of type $s$ in zone $k$ over the performance period $\pi$.

Similarly the rate at which vehicle-hours are performed in period $\pi$ on links of type $s$ within zone $k$ can be defined:

$$VH_p^{\pi sk} = \left( \sum_{t=t_\pi}^{t_{\pi+1}} \sum_{\forall l \varepsilon L_{sk}} \sum_{\forall n \varepsilon V_l(t)} h_n(t) \right) / H_\pi . \tag{2}$$

Then performed speed, $U_p$, is given by:

$$U_p = VK_p^{\pi sk} / VH_p^{\pi sk} . \tag{3}$$

Plotting $U_p$ against $VK_p$ for different levels of underlying demand provides the basic performance relationship between speed and flow.

**The tracking approach**

The tracking approach is used to record the time spent and distance travelled by each vehicle generated in a generating period $g$. Vehicles are tracked until they reach their destination or, in the case of highly congested conditions, until the end of the simulation, which continues for an hour after the end of the last generating period.

The following definitions will be used to describe the data collected:

$g$ the generating period during which vehicles enter, or queue to enter, the network;

$H_g$ the length of the generating period;

$\gamma$ the adjusted generating period;

$H_\gamma$ the length of the adjusted generating period, taken as the average of the generating period and the time between exit of the first and last vehicles generated;

$t_g$ the start of the generating period that will end at $t_{g+1} = t_g + H_g$;

$n_g$ an index for a vehicle generated in period $g$;

$d_{ng}(t)$ distance travelled by vehicle $n$ generated in period $g$ from time $t - \lambda$ to time $t$;

$h_{ng}(t)$ $\lambda$ (= 1 second) if vehicle $n$ from period $g$ is present at time $t$, otherwise zero;

$t_{end}$ the end of the simulation;

$F$ a factor by which the free flow demand level is increased in subsequent simulations.

$H_\gamma$ is calculated as the average $H_g$ and the time elapsed between the exit times for the first and last vehicle generated. This is an approximation, but does not substantially affect comparison between the results presented below.

For supplied flow, the vehicle-km actually travelled are collected for each vehicle as it travels through the network. The vehicle-km supplied for vehicles generated in period $g$ on links of type $s$ within zone $k$ can be defined as:

$$VK_s^{gsk} = \left( \sum_{t=t_g}^{t_{end}} \sum_{\forall l \varepsilon L_{sk}} \sum_{n_g} d_{n_g}(t) \right) / H_\gamma \quad . \tag{4}$$

We have summed the distance travelled at each time step $t$ by vehicles generated in period $g$ and present on links of type $s$ in zone $k$ over the period $t_g$ to $t_{end}$.

The demand flow from period g for link type $s$ in zone $k$ is calculated for any value of $F$ as the vehicle-km supplied under free-flow conditions (that is, on free-flow routes), factored by the appropriate value of $F$. That is:

$$VK_d^{Fgsk} = F.VK_s^{Fgsk}|_{F=1}. \quad (5)$$

For supplied time, the vehicle-hours are again collected for each vehicle as it travels through the network. The vehicle-hours supplied for vehicles generated in period $g$ on links of type $s$ within zone $k$ can be defined as:

$$VH_s^{gsk} = \left( \sum_{t=t_g}^{t_{end}} \sum_{\forall l \varepsilon L_{sk}} \sum_{\forall n_g \varepsilon V_1(t)} h_{n_g}(t) \right) / H_\gamma \ . \quad (6)$$

Note that the vehicle-hours are defined to include the external queuing time for entry links.

Then supplied time per km, $T_s$, is given by:

$$T_s = VH_s^{gsk}/VK_d^{gsk}. \quad (7)$$

Plotting $T_s$ against $VK_d^{gsk}$ for different levels of underlying demand provides the basic supply curve between time per km and demand flow.

## Sample Results

### The range of conditions tested

Initial analysis was based on hypothetical grid and ring radial networks, which were used with a range of matrix structures to assess the effect of different distributions of movement on two very different network configurations. The results for the speed-flow relationship and the three curves of Figure 5 are presented below. Subsequent analysis used a single link, a single loop, and a single intersection to investigate issues of spatial and temporal dependency, as outlined subsequently. The final stage of the project applied the analysis methodology to a SATURN network of Cambridge; the results of this stage are presented elsewhere (May and Shepherd, 1996) and summarised briefly in the next section.

**Figure 6**
*Performed Speed vs Performed Travel*
*6 x 6 Grid Network : Total Network*

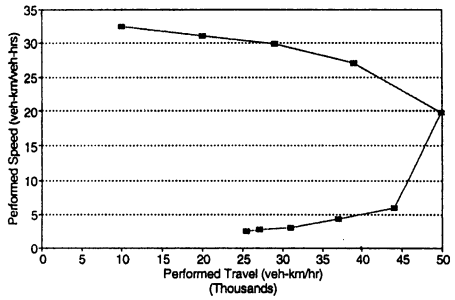
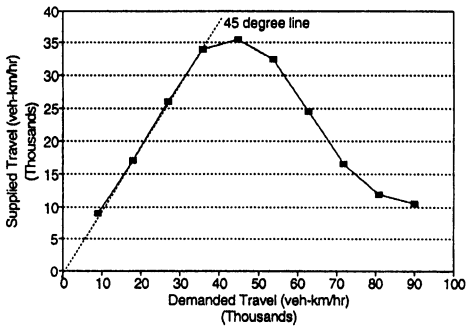



**Figure 7**
*Supplied Flow vs Demanded Travel*
*6 x 6 Grid Network ODA : Total Network*




**6 × 6 grid network**

In this section we present the four key relationships in Figures 2 and 5 for a hypothetical 6 × 6 grid network consisting of 36 signalised junctions with 24 external origins/destinations and 5 internal origin/destinations. Traffic entering the network queues at signalised junctions at the end of the entry links. A fixed matrix pattern, referred to as ODA, was used throughout, together with a fixed demand profile split into four 15 minute periods. Each point in the figures pre-

sented here relates to a single level of demand, represented by the factor $F$, and has been obtained from a separate simulation. Each demand level involved a uniform trip rate being applied for an hour, with analysis based on the middle 30-minute period, to allow the network to become loaded in the first 15-minute period, and vehicles still to be tracked in the final 15 minutes. Such a demand profile is, of course, unrealistic, but has been used to simplify the interpretation of the results.

Figure 6 shows the simulated speed-flow relationship averaged over the total grid network, plotting the speed performed ($U_p$) against the flow performed ($VK_p$). Each point represents the results from one simulation with demand increasing (increasing $F$) as one moves clockwise round the curve and is calculated from time-slice analysis. As demand is increased the physical performance of the network reaches a maximum ("capacity") beyond which any further increases in demand result in a drop in actual flow performed and a drop in speed. This particular network shows a classic bending-back curve as was suggested in Figure 2. While the first four points reflect a reasonably linear relationship, it is clear that a linear speed-flow relationship, as used in recent studies (Bates *et al.*, 1991; Oldfield, 1993), is not an appropriate performance curve for this network.

Figure 7 shows the supplied flow ($VK_s$) versus demanded flow ($VK_d$). It should be noted (see above) that at higher demand levels, supplied flow is measured over a longer period than demanded flow. For demands up to the fourth level, all demand is satisfied. Beyond this level, supplied flow soon starts to fall as congestion reduces network performance. This replicates the hypothesised shape in the top part of Figure 5, though there is no evidence of re-routeing.

Figure 8 shows the simulated supply curve, plotting the supplied time per km ($T_s$) against demanded flow ($VK_d$). The curve is relatively flat as long as demand is below capacity, but rises quickly as capacity is exceeded, in line with the hypothesised shape in the bottom part of Figure 5. Were demand determined solely by travel time, the intersection of this relationship and a demand curve for the network would determine the equilibrium vehicle-km that would actually be demanded.

Figure 9 shows the relationship between pseudo-speed and demanded flow. This is the inverse of Figure 8 and, as noted above, includes the effect of queuing to enter the network. As argued earlier, it is relationships such as this, rather than the performance curve in Figure 6, that should be used to represent supply in area models. By comparison with Figure 6, it can be seen that the first four points (up to about 80 per cent of "capacity") reproduce the performance speeds almost exactly, and the shape for this network is reasonably linear, corresponding to the linear relationships used in models such as that of Evans (1992) and MVA's START model.

**Figure 8**
*Supplied Time/km vs Demanded Travel*
*6 x 6 Grid Network ODA : Total Network*

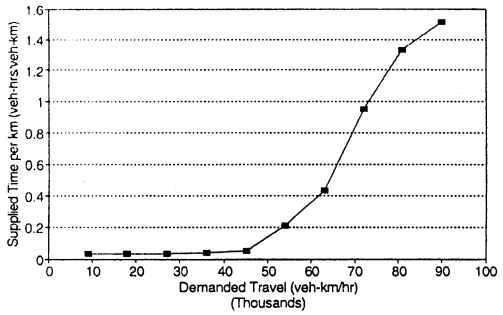
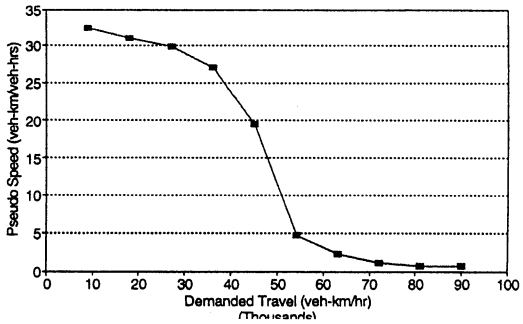



**Figure 9**
*Supplied (pseudo) Speed vs Demanded Travel*
*6 x 6 Grid Network ODA : Total Network*




However in the subsequent points (a) the pseudo-speed is lower than the performance speed (reflecting partly the queuing to enter the network), and (b)the relationship becomes markedly non-linear. This indicates the potentially serious problems of extrapolating observed speed-flow relationships up to and beyond "capacity".

Experiments were also made to test the sensitivity of the results to the shape of the matrices. Matrices were formed by doubling the demand between certain OD pairs in the ODA Matrix as follows:

Matrix B : heavy inbound
Matrix C : heavy outbound
Matrix D : heavy clockwise
Matrix E : heavy anti-clockwise.

It was found that the grid network supply curve is insensitive to such changes in OD pattern. This is probably due to the similar routeing opportunities available in a grid network.

**Ring-radial network**

The ring-radial network consisted of eight radials and three rings with eight external, eight internal, and eight central origins/destinations. The network was split into nine zones: one central zone, four identical inner zones, and four identical outer zones. The demand levels were generated in the same way as for the grid network. Figures 10 to13 show comparable results to Figures 6 to 9 for the ring-radial network with a similarly configured matrix.

**Figure 10**

*Performed Speed vs Performed Travel*
*Ring-Radial Network ODA : Total Network*

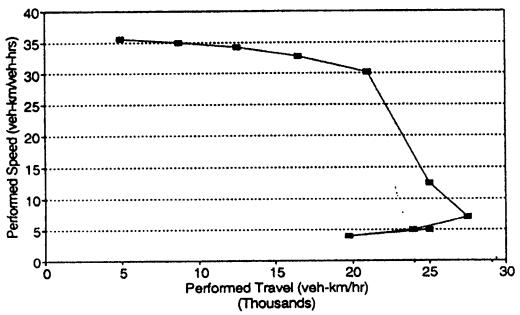



**Figure 11**

*Supplied Travel vs Demanded Travel*
*Ring-Radial Network ODA : Total Network*

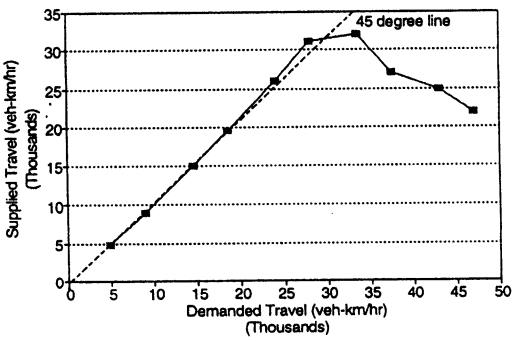


**Figure 12**
*Supplied Time/km vs Demanded Travel*
*Ring-Radial Network ODA : Total Network*

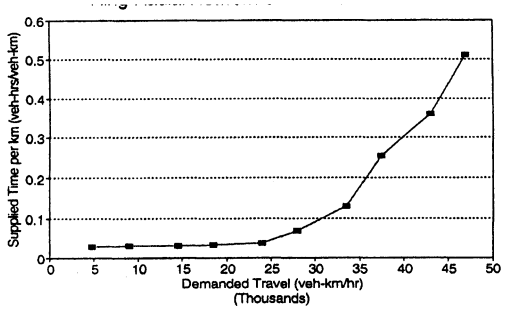
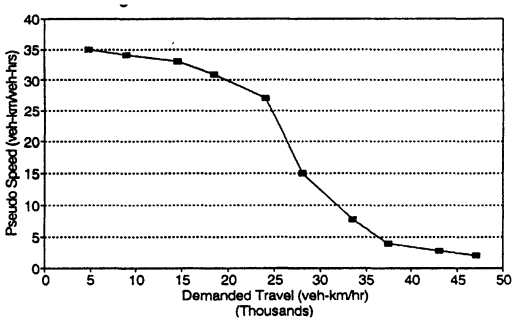



**Figure 13**
*Supplied (pseudo) Speed vs Demanded Travel*
*Ring-Radial Network ODA : Total Network*




Once again, the performance curve, (Figure 10) plotting speed performed ($U_p$) against flow performed ($VK_p$) shows a classic bending-back curve, with "capacity" reached at demand level 7.

Figure 11 compares supplied flow ($VK_s$) with demanded flow ($VK_d$). Here the full demand is satisfied for the first six demand levels, but this time with evidence of re-routeing. However, by "capacity" (demand level 7) not all demanded flow is being performed.

Figure 12 shows the simulated supply curve, plotting supplied time per km ($T_s$) against demanded flow ($VK_d$). Here supplied time per km remains at a low level until demand point 5, but has risen significantly by the time that "capacity" (demand level 7) is reached.

Figure 13 shows the relationship between "pseudo-speed" (the inverse of $T_s$) and $VK_d$. Here the first five points are noticeably linear, but the curve, as in Figure 9, then becomes markedly non-linear. An assumption of a continued linear shape in this area, when translated to the supply curve (Figure 12) would substantially affect the estimation of equilibrium flows close to capacity.

**Figure 14**
*Performed Speed vs Performed Travel*
*Ring-Radial Network ODA : Total Network*

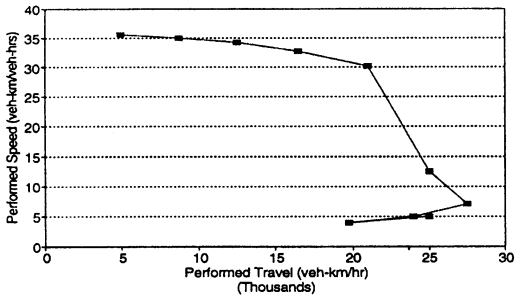
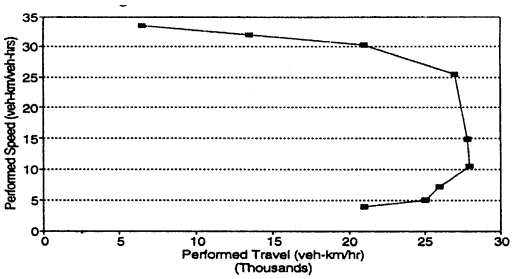



**Figure 15**
*Performed Speed vs Performed Travel*
*Ring-Radial Network ODB : Total Network*




Figures 14 to 17 compare the performance curves for the base matrix A (used in Figures 10 to 13) and a heavy inbound matrix B for the total network and for the inbound links of one of the outer zones, zone 6. (Figure 14 is of course identical to Figure 10.) There is a major difference in the performance of the network for matrices A and B. Matrix B (Figures 15 and 17) gives a high-

er maximum flow indicating a higher capacity. This is due to the higher inbound demand and the complex interactions between the links within the network. A heavy outbound and a heavy through matrix were also simulated and also generated results that were different from the base matrix and from one another. The different distributions of demand between directions results in different allocations of flow between competing links at junctions, and hence different patterns of congestion. These in turn lead to differences in aggregate performance. This has been termed spatial dependency and is explored further below.

**Figure 16**
*Performed Speed vs Performed Travel*
*Ring-Radial Network ODA Zone 6 Inbound*

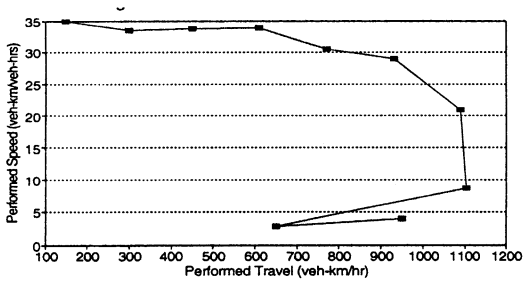
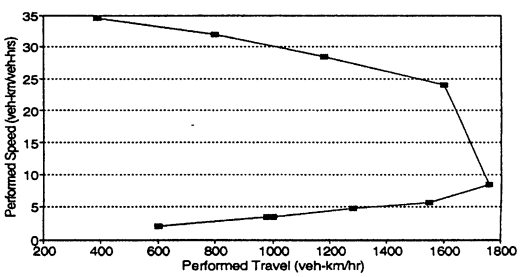



**Figure 17**
*Performed Speed vs Performed Travel*
*Ring-Radial Network ODB Zone 6 Inbound*



**Simpler networks**

Simpler networks have been used to investigate some of the underlying processes, and particularly temporal and spatial dependency, in more detail. May and Shepherd (1995) describe the results from simulations of single link and loop networks. While the single link represents a simple linear transition from A to B, the single loop involves a route from A to B that intersects with itself at a junction. The single link with a free-flowing exit has a performance curve with a free-flowing regime, a lower speed regime, and finally a fixed point when the physical link is full and the traffic is exiting at a constant rate (capacity). In contrast the single loop shows the classic bending-back of the performance curve as the loop becomes blocked and gridlock forms. Here capacity may be defined by the highest sustainable flow; any further increase in demand causes the performance to deteriorate and move to the lower part of the performance curve.

The single link and loop networks were also used to show the effect of build-up of congestion or temporal dependency. Each measure was aggregated according to the generating sub-period (each sub-period, or slice, being 15 minutes). This showed the temporal dependency of the data collection method as the curves diverged beyond the free-flow regime as shown for the single loop in Figure 18. Each slice is, as would be expected, dependent upon the previous slice, and therefore also dependent on the previous demand level and how long that demand level has been sustained. In the case of the single loop the congestion builds up until, by the last slice, gridlock occurs.

Shepherd (1995) describes the use of a single priority junction network, in which major route AB intersects with minor route CD, with traffic on CD giving way to that on AB, to investigate the effect of spatial dependency. Two separate OD patterns were simulated, the first with equal flows on the major and minor routes, the second with double the flow on the major route. Figure 19 gives an example of speed versus demanded flow curves for the minor (give way) route for both OD patterns. The curves are significantly different from one another even though the flow demanded for this route is the same for both OD patterns at each corresponding point. The minor route is affected by the flow on the major route as it must give way. As the flow on the major route is doubled then the capacity of the minor route is reduced, resulting in a shift in the curve for the same demanded flow. The performance curve for the minor route is a classic bending-back curve and has changed due to a change in flow on another route. These results for single links feed through in the simulation of larger networks to produce the shifts in area speed flow relationships observed in Figures 14 to 17. In such networks, changes in the distribution of flows, and hence in performance curves, can be induced both by changes in OD patterns and by reassignment.

**Figure 18**
*Supplied Time/km vs Demanded Travel*
*Single Loop : Consecutive Time Slices*

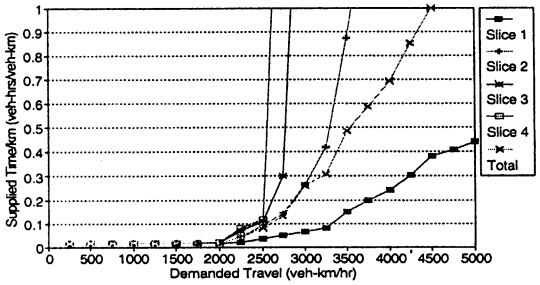
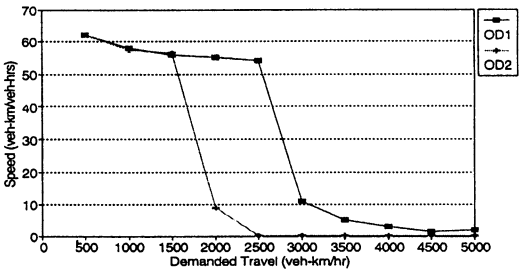



**Figure 19**
*Supplied Speed vs Demanded Travel*
*Single Junction : Minor Link*




## Comparison with other work

The curves produced by microsimulation for the grid and ring-radial networks can be compared with previous attempts at building area speed-flow relationships using either real data or simulated data. However, none of the literature reviewed has identified the need to distinguish between performance and supply curves. Ardekani and Herman (1987) used aerial photography and ground data in combination to produce a "performed" speed density (concentration) re-

lationship for Austin and Dallas, Texas. Various relationships were hypothesised for this performance curve, none of which exhibited the classic backward-bending portion of the curve as in Figures 6 and 10. As with most practical observations the data range was limited to speeds between 20 and 25 km/h (for flows below capacity), which by definition lie on the upper part of the performance curve. While the relationships generated appear superficially to be similar to Figures 9 and 13, this is simply a result of the extrapolation process adopted.

Ohta and Harata (1989) produce an aggregate performed speed flow curve by simulation that is in practice similar in shape to Figures 9 and 13. However, this results from an assumption that there is a constant speed of 10 km/h for flows above capacity at the individual link level. Indeed the assumptions about the model are critical in defining the aggregate curve; and as there are no capacity effects or interactions between flows on different links, then the supplied curve will be equivalent to the performed curve for this particular model.

Harrison *et al.* (1986) used observed speed flow relationships at the link level that, when aggregated, give a curve similar in shape to Figures 9 and 13. However, the range of observed data was again limited, and assumptions were made about the shape of the curve at lower speeds. For speeds less than 5 km/h the journey time was assumed to increase linearly to represent queuing time. This assumption produces the similarity to the curves in Figures 9 and 13, and comes closest to our approach, in that it allows for time spent queuing. However, as with other relationships in the literature, it fails to address fully the requirements of a supply curve.

## Implications

### Network performance curves

The results presented here suggest that networks, like links, can exhibit the classic backward-bending relationship between speed performed and flow performed (measured here in veh-km/h) that is evident on individual links. As with individual links, a bottleneck is needed if the lower part of the curve is to be generated. However, as explained in the second section, performance curves of this kind are inappropriate for use in modelling the interaction between demand and supply. For this purpose, speed needs to be related to the flow demanded rather than to the flow that can be performed in practice.

**Network supply curves**

The required form of network supply curve is one that gives supplied time/km against flow demanded (again measured in veh-km/h): the time needs to include the effect of queuing. The intersection of such a curve with a demand curve (after allowance for other components of "generalised cost") can be used to determine the flow that actually occurs.

In many current applications such a curve is being implied by using a speed-flow relationship. The analysis here shows that the corresponding inverse supply relationship, between speed supplied (here measured as a pseudo speed, because of queuing effects) and flow demanded can be generated at a network level. However, there is no evidence to support the widely held assumption that this relationship is linear throughout the full range of demand.

**Changes in OD pattern**

The ring-radial network used above showed that the performance and supply curves obtained were sensitive to the overall demand pattern. Lower supplied time per km ($T_s$) was obtained for a given demanded flow with the heavy inbound matrix, apparently because of the ease with which minor movements could re-route to accommodate major ones. However, for the grid network the performance and supply curves appear insensitive to even quite substantial changes in the distribution of trips. The effect is probably less for the grid network because the grid structure provides a wide range of similar alternatives, while the radial and orbital routeing options in a ring-radial network are inevitably different. This issue of spatial dependency was explored further in the previous section. Changes in the performance and supply curves can be induced both by changes in OD patterns and by reassignment, and the latter can result directly from certain policies such as those that restrict through traffic.

**Temporal dependency**

The supply curves are sensitive to the temporal distribution of demand and in particular to delays inherited from earlier time periods, as illustrated in the previous section. Thus the costs incurred by a particular demand in a given time period will be determined in part by the demand in earlier time periods. While this process is obvious, it presents very considerable difficulties for the modelling of network supply. The current strategic modelling approach is to assume a constant demand during each time period, that no single time period will be over-capacity, that there is thus no need to pass queues from one period to the next, and that the time periods can be considered to be independent. Such assumptions may be acceptable if the whole peak period, including its shoulders, is modelled as one generating period with a constant demand given by the average for that period, since all demand should be satisfied over that period.

Even so, it could be expected that total travel time in a peak period with a given level of demand would be higher if the demand was peaked than it would if it was evenly distributed. Further research into this area is required.

**Implications for strategic modelling**

The study has confirmed that it is possible to generate area speed-flow relationships from simulation models. However, the relationship required for strategic modelling is between speed supplied and flow demanded, and cannot therefore be observed empirically. There is no evidence from the study to indicate that the linear relationship between speed and flow assumed in many current applications applies over the full range of demand. This may represent a problem if the equilibrium point is beyond the linear section.

The relationships produced are dependent on network characteristics and capacities and, for certain network configurations, on the shape of the matrix. This implies, in theory, that new relationships will be needed if the matrix shape, network capacity or assignment is significantly affected by the strategy tested, such as restraint in a city centre or a new radial road. Subsequent research for a new strategic model for Manchester has confirmed this result.

Supply curves will also be sensitive to the temporal distribution of demand. It is difficult to see how such interactions can be incorporated in a strategic model; current practice is to assume that the demand is constant over a long period and that all demand is supplied within such a period. This assumption is acceptable so long as no sub-period within the peak period is over capacity. Once a sub-period becomes over capacity then the results of a flat average demand profile will underestimate the true peak profile congestion and the shoulders of the peak will also be affected. As noted above, further research into this issue is required.

Subsequent research used a SATURN network of Cambridge to collect area speed-flow measures and to use them to develop a strategic network of Cambridge. SATURN (van Vliet, 1982) is a widely used combined assignment and simulation model for the analysis for urban networks. The results confirmed that area speed flow relationships and supply curves could be defined and collected from a SATURN network. However, their use in an aggregate model did not produce a satisfactory representation of the full network results, with very different network conditions in the base ("do-nothing") case (May and Shepherd, 1996). The area speed flow relationships were insensitive to a particular change in OD pattern, (formed by applying road pricing, giving less centre-bound traffic), which is in line with the grid network rather than the ring-radial network results obtained using NEMIS. However, when the full and aggregate networks were used separately to test the effects of road pricing in Cambridge, very different results were obtained. Because road pricing induces both re-

routeing and changes in the shape of the matrix, it gives rise to spatial dependency impacts of the kind described in the previous section. These could not readily be represented in the area speed-flow relationships, which thus gave different results. This has important implications for the use of aggregate networks in the strategic modelling of transport policies.

**Potential for further research**

The methodology developed in this study has the potential for extension in a number of ways. One of the most important requirements is to understand the issue of temporal dependency more fully, and to extend it to investigate the impact of peaked demand profiles. Further investigation of the causes of spatial dependency in different types of network would also be worthwhile. Beyond these, further study is needed of the impact of different policies, and particularly those that change the shape of the matrix, and increase or reduce trip length. These could usefully be studied on both hypothetical and real networks.

# References

Ardekani, S., and R. Herman (1987): "Urban Network-wide Traffic Variables and their Relations". *Transportation Science*, 21, 1-16.

Bates J. J., M. Brewer, P. Hanson, D. McDonald, and D. Simmonds (1991): "Building a Strategic Model for Edinburgh". Paper presented at the 19th PTRC Summer Annual Meeting, Brighton. London: PTRC.

Department of Transport (1996): *Design Manual for Roads and Bridges.* Vol 13, Sec 1, Part 5, Ch 9. London: DoT.

Duncan, N. C., A. W. Christie, and M. Marlow (1980): "Traffic Speeds in Towns: Further Analysis of the Urban Congestion Surveys". *Traffic Engineering and Control*, 21, 576-9.

Evans, A.W. (1992): "Road Congestion Pricing: When is it a Good Policy?" *Journal of Transport Economics and Policy*, 26, 213-44.

Greenshields, B. (1934): "A Study of Traffic Capacity". Paper presented at the 14th Annual Meeting of the Highway Research Board, Washington, DC.

Hall F. L., and F. O. Montgomery (1993): "The Investigation of an Alternative Interpretation of the Speed-Flow Relationship for UK Motorways". *Traffic Engineering and Control*, 34, 420-25.

Harrison, W. J., C. Pell, P. M. Jones, and H. Ashton (1986): "Some Advances in Model Design Developed for the Practical Assessment of Road Pricing in Hong Kong". *Transportation Research*, 20A,135-44.

Hills, P. J. (1993): "Road Congestion Pricing: When is it a Good Policy? A Comment". *Journal of Transport Economics and Policy*, 27, 91-9.

Mauro, V. (1991): "Evaluation of Dynamic Network Control: Simulation Results using NEMIS Urban Microsimulator". Paper presented at the Transportation Research Board Annual Meeting, Washington DC.

May, A. D., and K. Gardner (1990): "Transport Policy for London in 2001: the Case for an Integrated Approach". *Transportation*, 16, 257-77.

May, A. D., and M. Roberts (1995): "The Design of Integrated Transport Strategies". *Transport Policy*, 2, 97-105.

May, A. D., and S. P. Shepherd (1995): "An Investigation of Area Speed Flow Relationships by Micro-simulation". Paper presented at 23rd European Transport Forum, Manchester. London: PTRC.

May, A. D., and S. P. Shepherd (1996): "Area Speed Flow Relationships and Network Aggregation". Paper presented at Second International Conference on Urban Transport and the Environment. *Urban Transport 96.* Barcelona: Computational mechanics publications.

Neuberger, H. (1971): "The Economics of Heavily Congested Roads". *Transportation Research*, 5, 283-93.

Ohta, K., and N. Harata (1989): "Properties of Aggregate Speed-flow Relationship for Road Network". *Transport Policy, Management and Technology towards 2001 : Selected proceedings of the Fifth World Conference on Transport Research, vol IV: Contemporary Developments in Transport Modelling.* Ventura, CA: Western Periodicals Co.

Oldfield, R. H. (1993): *A Strategic Transport Model for the London Area.* Research Report 376, Transport Research Laboratory, Crowthorne.

Shepherd, S. P. (1995): "Area Speed Flow Relationships: The Effects of Link Dependency and Reassignment using Two Link Networks". Working Paper 448, Institute for Transport Studies, University of Leeds.

Small, K. A. (1992): *Urban Transportation Economics*. Harwood Academic Publishers.

Small K. A., and X. Chu (1997): "Hypercongestion". Paper presented at the Annual Meeting of the American Real Estate and Urban Economics Association, New Orleans.

Thomson J. M. (1967): "Speeds and Flow of Traffic in Central London 2: Speed/Flow Relationships". *Traffic Engineering and Control*, 8, 721-25.

van Vliet, D. (1982): "SATURN : a Modern Assignment Model". *Traffic Engineering and Control*, 23, 578-83.

Walters, A. A. (1961): "The Theory and Measurement of Private and Social Cost of Highway Congestion". *Econometrica*, 29, 676-99.

**Published paper**

A. D. May, S. P. Shepherd, and J. J. Bates (2000) Supply Curves for Urban Road Networks. Journal of Transport Economics and Policy, 34(3) 261-290