# CHAPTER 2

# A LITERATURE REVIEW OF THE SPEED-FLOW-DENSITY ECONOMIC MODELLING OF ROAD CONGESTION

## 2.1 Introduction

It has long been recognized that road traffic congestion can cause an economic externality in urban areas and on trunk roads (see, for example Pigou, 1920, and Nash, 2007). Many economists from Pigou (1920) onwards have analysed this problem. It is argued in this review that there are limitations in past analyses and that some of these stem from the emphasis of previous analyses on flow as the basic cause of congestion rather than recognising that flow, speed and density are determined simultaneously. It is clear that there are many debates in the road congestion literature and the models developed in the following chapters attempt to examine some of these sometimes difficult to follow disputes.

This Chapter is in ten further sections. The next section considers the policy implications of the marginal external congestion time cost. The third section provides a brief history of the economic analysis of road congestion up to and including the work of Vickrey (1969). The fourth section provides a presentation of the standard view of the speed-flow-density model of road congestion. The fifth section considers empirical studies investigating the marginal external cost of ordinary congestion are examined. In the sixth section, dynamic models of congestion are examined and the importance of multiperiod analysis is emphasised. The seventh section develops a model of road congestion using

density, i.e. road occupation. The eighth section examines the literature on the economics of hypercongestion. The new model developed in the seventh section is used to help explain the debates between Else (1981 and 1982) and Nash (1982); Demza and Gould (1987) and Evans (1992a); Evans (1992a and 1992b) and Hills (1993); and Ohta (2001a and 2001b) and Verhoef (2001). In the ninth, studies that have considered the impact of different types of vehicles on road congestion are examined. The tenth section considers the importance of the value of time in estimating the marginal external congestion time cost and for setting the price in road congestion charging policies. Finally, the concluding section suggests how models of the economics of road congestion can be improved.

## 2.2 Policy Implication

This thesis aims to estimate the marginal external time congestion cost which is important for constructing an efficient road congestion pricing policy.[1] Therefore, this section will discuss the importance of road congestion pricing policy and the implementation of such policies.

It has been widely agreed that road congestion causes a huge economic impact, especially in urban areas and major highways (e.g. Newbery 1987; Niskanen and Nash, 2008; Mattsson, 2008; De Palma and Lindsey, 2011 and Lindsey and Verhoef, 2000b). Time lost in travelling is clearly seen as an important direct cost (De Palma and Lindsey, 2011). Whereas, the uncertainty of travel time reliability is a significant indirect cost. It implicitly causes the inconvenience of trip rescheduling or changing to alternative travel modes, and it possibly results in greater fuel

[1] It is important to make clear that the analysis of road externality discussed in this thesis ignores other externalities.

consumption and vehicle wear and tear. Moreover, in the long run, the uncertainty will lead to costs of relocating residences and jobs (Lindsey and Verhoef, 2000b).

As a policy response to solve this problem, in terms of economic efficiency, it is often proposed to charge road users the marginal external congestion cost they impose (Mattsson, 2008). For example, Goodwin (1995, p. 149) suggests that "charging people something for the external costs of congestion and environmental damage will be more efficient than not charging them." Likewise, De Palma and Lindsey (2011, p. 1) explain that "Congestion pricing has a big advantage over other travel demand management policies in that it encourages individuals and firms to adjust all aspects of their behaviour: number of trips, destination, mode of transport, time of day, route and so on, as well as their long-run decisions on where to live, work and set up business."

Thus, the review of congestion pricing policies is important. Before moving on to a basic analysis of the economic efficiency, it is emphasised that this analysis assumes all other necessary conditions hold to ensure a first-best outcome. Thus, for example, there are no other conflicting externalities, no transaction costs (Proost, 2011) and perfect information for every road user. Figure 2.2.1 is the basic underlying diagram (see for example, Rouwendal and Verhoef, 2006, p. 107). The horizontal axis represents flow quantity (vehicle-km) and the vertical axis shows the unit cost of travelling on a uniform and homogeneous road segment (£/veh-km). The curve APC[2] is the average private cost, which is the product of value of time and the travel time for a vehicle-kilometre on the road

[2] APC under this analysis does not include the vehicle operating costs and the value of time is assumed constant for all road users.

segment. The demand curve D represents the demand for travelling on

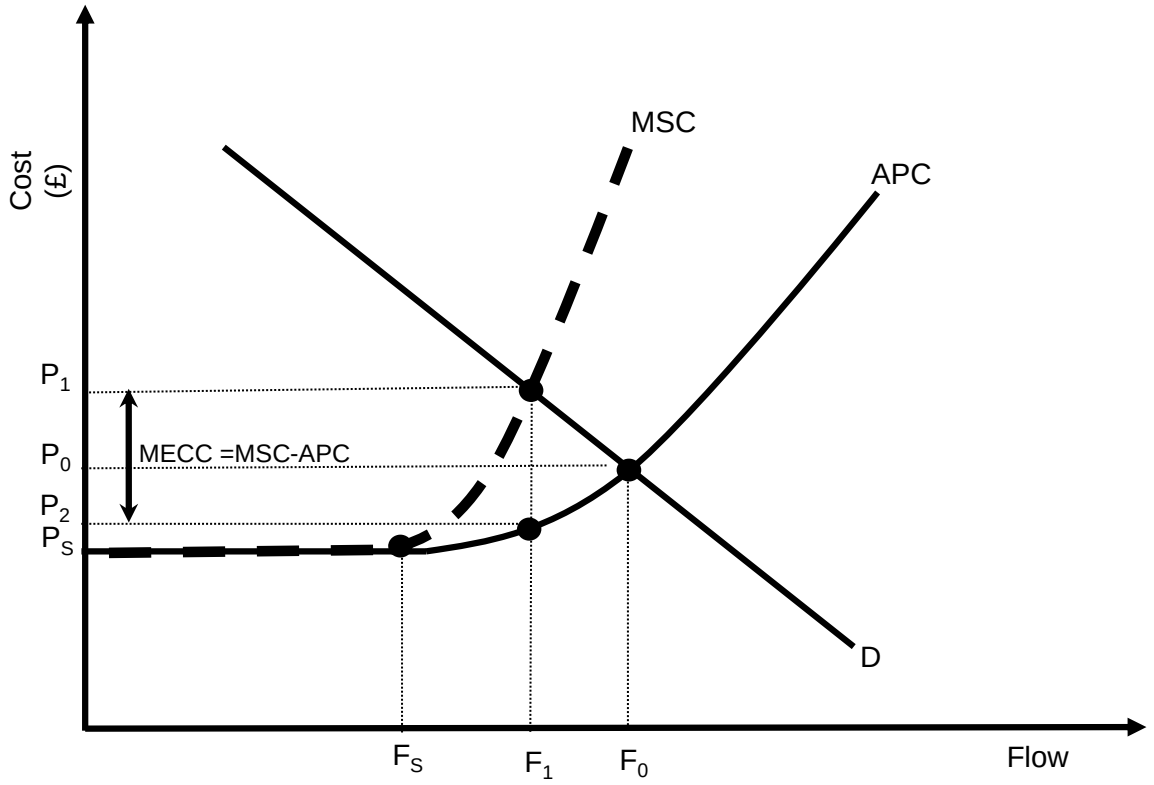


**Figure 2.2.1 The Congestion Pricing Analysis**

this road segment. At the beginning (the flow range between 0 and $F_s$), APC and the marginal social cost MSC are equal and constant at $P_s$. It is important to note that every vehicle pays the same cost. After the flow reaches the level $F_s$, the road becomes congested. As a result, both APC and MSC increase but MSC is steeper than APC. The difference between MSC and APC is the marginal external congestion cost MECC. The value of MECC is measured as the product of the value of time and the delay time caused by an additional vehicle-kilometre. For the situation without congestion pricing, the market outcome is where the demand and average private cost schedules intersect at D=APC, at $P_0$ and $F_0$. In order to correct the market distortion, every vehicle should be responsible for the cost it imposes on others. For this reason, the optimal condition is at where D=MSC,

taking into account the MECC and the price is higher at $P_1$ and the optimal flow level is lower at $F_1$. This Figure is an oversimplification as it assumes a one period model with complete uniformity of conditions. In reality, flows and speeds vary over the day and greatly complicate theoretically correct congestion analysis.

Though the congestion pricing concept outlined above is attractive as a first-best[3] solution, it is not possible in practice. But it provides a useful basic economic starting point for implementing the policy (Small and Verhoef, 2007; Santos and Verhoef, 2011 and Proost, 2011). However, economists understand the road congestion in reality is dynamic with congestion changing over time and place and is caused and experienced by heterogeneous road users (Santos and Verhoef, 2011). For this reason, the impact of such heterogeneous circumstances on marginal external congestion time costs is examined[4]. In no particular order and not exhaustively, there are differences in: on and off-peak congestion (see Vickrey, 1963; Lindsey and Verhoef, 2000b); the values of time (see Van Den Berg and Verhoef, 2011); drivers' behaviour; types of vehicle; and road characteristics in the same network ( see Peirson, Sharp and Vickerman, 2001; De Palma and Lindsey, 2002a and Rouwendal and Verhoef, 2004). In addition, there are market distortions from other related economic sectors that also affect a policy of marginal social cost pricing. For instance, transport related locational decisions such as land use or general economic development (Nash and Sansom, 2001)

[3] First-best situation is the situation when the equality between prices and marginal costs (Rouwendal and Verhoef, 2006, p. 108). In addition, Small and Verhoef (2007, p. 137) define that "The rules for marginal-cost pricing discussed in the previous section are often referred to as "first-best" because there are no constraints on the pricing instrument and there are no market distortions other than the congestion externality.

[4] Ignoring heterogeneity in estimating the first-best can result in negative, zero, or positive biases, depending on the type and extent of heterogeneity (Van Den Berg and Verhoef, 2011, p. 60). For instance, after taking into account of the heterogeneity of value of time with the equally revenue redistribution condition, the low value of time users face greater loss than that which is estimated in the first-best (Arnott et al, 1994).

have to be related to the second best efficient price in congestion charging. A more general distortion is government's and road users' lack of access to and use of perfect information about congestion and time varying congestion charging (Santos and Verhoef, 2011 and Proost, 2011). These concerns lead to a question of transparency in charging systems and, for example, Santos (2004b, p. 316) worries that "drivers would not know the congestion charge they would be required to pay before starting their journey." Importantly, these concerns lead to a question of transparency in charging systems and with the "behavourial" response of road users may make measurement and implementation of an optimal congestion pricing of road network difficult (Newbery and Santos, 1999, Santos, 2004b and Small and Verhoef, 2007).[5]

Santos et al., 2010, lists a number of problems with first best charging policies. It may be difficult to measure the marginal external congestion cost. It is possible that other existing transport policies may even lead to overcharging for congestion externalities, eg the subsidising of other modes of transport. Policies in other related sectors, eg vehicle insurance (see Peirson, Skinner and Vickerman (1998) and income taxes (see Verhoef, 2000) could also affect the efficiency of congestion charging.

However, failure in achieving the efficient allocation of road space may not only be caused by technical and administrative problems but also through the public acceptability of such schemes. While many economists see social benefit in

[5] First-best is unable to provide explanations for these questions e.g. how to set tolls, how to cover common costs, what to do with any excess revenues, whether and how "losers" from tolling previously free roads should be compensated, whether to privatize highways and how to set the road pricing dealing with the political opposition or to compensate the major losers (Lindsey, 2009 and Santos,2000).

congestion pricing policies, but the introduction of this policy is often unsuccessful (Jones, 1998). Harrington (2001) and Small (1992a) suggest that it is not easy to convince people to pay for something which was previously free. If road congestion pricing policy is to be introduced through a democratic process, it is necessary to convince voters of the net benefits of charging for the use of roads (see for example, Schade and Schlag, 2003; King et al, 2007 and Hårsman, and Quigley, 2010). With this consideration, there are different perspectives between economists and politicians. Economists often see congestion pricing policy as giving an efficient road space allocation via the market. However, politicians are generally suspicious of charging policies as the general public sees it as an additional cost (Oberholzer-Gee and Wech-Hannermann, 2002 and King et al., 2007). Most politicians supply the policies demanded by voters in exchange for their votes. For this reason, the congestion pricing policy is not necessarily politically popular. Nonetheless, there is an exceptional example. Ken Linvingstone, a British politician proposed the congestion charging as a major part of his political campaign[6] for the election of the London Mayor in 2000 and he won with 58% of votes cast (Peirson and Vickerman, 2008). However, winning the election was just a first step[7]. Success in implementation of charging policies was

[6] Ken Livingstones, the Mayor of London delivered an inspiring speech: "it is bound to be messy, it is bound to involve mistakes, but in the end everyone will look back and agree it was a good idea and they won't be able to recall why they ever resisted the change….Yet changing transport patterns that have built up over 30 years will clearly take time and lashings of criticism." (Livingstones, 1998, is quoted in Peirson and Vickerman, 2008, p. 81.).

[7] "The primary political challenge for congestion pricing is thus not to maximize the number of winners, but rather to overcome initial antagonism to the idea. Once pricing becomes the status quo, its political problems will steadily diminish because it will benefit from the same political inertia that now works against it." (King et al., 2007, p. 114).

**Table 2.2.1 Second-Best Pricing Policies**

| No. | Second-Best Pricing | Issues |
|---|---|---|
| 1. | Priced and un-priced lanes in parallel | Departure time rescheduling with heterogeneous value of time road users. |
| 2. | Different pricing at different streets in the same network | Traffic variation over different locations in the same network, including land-use effects in the broader urban structure. |
| 3. | Time-based pricing | Traffic variation over time |
| 4. | Distance-based pricing | For discouraging short distance trips |
| 5. | Area-based (per entry/exit) pricing | Bottleneck congestion (variation over different locations) |
| 6. | Earmarking revenues (e.g. for subsidizing the public transportation, expanding road capacity) | Heterogeneous value of time road users and taking into account of tax-distorted labour market, revenue distribution based on public acceptability |
| 7. | Parking Fees | Mobility behaviour in urban areas |
| 8. | Car pool or High Occupancy Vehicle Lane pricing | Non-identical road users with different demand elasticities on alternative transport choices |
| 9. | Stepwise pricing | Traffic variation over time and vehicle types |
| 10. | Capacity choice in relation to road pricing | Self-financing (for multiple local governments), private ownerships |

another challenge[8]. For this reason, the design of policy has to directly consider its political feasibility, its fairness and the nature of the resulting transport system (Small, 1992b). This also involves explicit consideration of the use of revenue streams[9] (Goodwin, 1997, p. 2). In other words, the revenue distribution is a key strategy to increase the support from local political leaders and groups who benefit from the congestion revenue (King et al, 2007, p. 112).

Because of the restrictions reviewed above, a second-best policy is often seen as more practical[10]. Each second-best policy should takes into account indirect effects, for instance, modes of transport (public or private), time of travelling (on/off-peak period), the parking cost, the resource reallocation between peak and off-peak periods, heterogeneous values of time, related financial policies, welfare distribution and political resistance.

Table 2.2.1 provides a list of suggested hypothetical second-best pricing policies and their related issues. The list includes two parallel routes with one unpriced route (Verhoef et al., 1995,1996; Braid, 1996 and Small and Yarn, 2001), different pricing at different streets in the same network (Small and Yarn, 2001 and Verhoef and Small, 2004), earmarking revenues (Mayeres and Proost, 2001 and Parry and Bento, 2001), time-based pricing (Ramjerdi et al., 2004), distance-based pricing

[8] For example, after winning the 2000 Mayor of London election, Ken Livingstones brought forward the congestion charging proposals. Accordingly, implementation was carried out with a consultation process with the public, local councils, business and representatives of road users. At the end, the implementation of the London Congestion Charging Scheme was successful with the integration of the improvements to the bus, underground and rail services, changes to the road network and a major overhaul of traffic management in London (Peirson and Vickerman, 2008, p. 80).

[9] For instance, the opinion poll reported by the Committee for Integrated Transport shows that Londoners increased their support (from 30% to 58%) to the London Congestion Charging Scheme after they were told that the revenue from congestion charging will be used for improving the public transportation service (Peirson and Vickerman, 2008, p. 80-81).

[10] Rouwendal and Verhoef (2006, p. 108) define second-best policy as the situation where the optimal price is determined in the context of market distortion.

(Jou et al., 2012), cordon-based pricing (Santos et al. 2001), carpooling and high-occupancy-vehicle lane (Yang and Huang, 1999), stepwise pricing of a bottleneck (Arnott et al.,1993a), parking fees (Glazer and Niskanen, 1992; De Borger, et al.,2001 and Calthrop et al., 2000), and capacity choice in relation to road pricing (Verhoef et al. 1996, De Borger et al., 2005 and Ubbels and Verhoef, 2008).

In conclusion, whether second-best policies trying to get close to a first-best are more effective in terms of increasing economic welfare is difficult to assess. Economic analysis should attempt to compare practical policies through simulation exercises as it is unlikely that any one second best policy can be analytically shown to be superior.

### 2.3 The First Studies of the Economics of Road Congestion

Adam Smith (1937) and Dupuit (1844 and 1849) were probably the first economists to consider the economics of road transport and, to a certain extent, congestion, see Lindsey (2006). Pigou (1920) was the first modern economist to consider the impact of tolls on congestion in a slightly confusing example of two different types of roads between the points A and D where differential taxation may "create an 'artificial' situation superior to the 'natural' one", (p. 194). Pigou is generally regarded as the discoverer of the concept of externalities though not using the term specifically, see Bohm (2008). As the quotation suggests, Pigou does not explicitly explain the causes of road congestion but hints at these causes. Knight (1924) developed Pigou's two road example making it clear that the roads differed in the possibility of congestion and considered whether private ownership and tolling would give an efficient outcome. However, neither Pigou nor Knight

considered how road congestion is generated and the importance of the concept of the marginal cost in dealing with the problem of congestion, see Lindsey (2006).

Vickrey was perhaps the first economist to consider explicitly the economics of road congestion and the underlying causes of congestion. Prior to the 1960s, it has been suggested that the causes of (not the solutions to) road congestion were mainly considered by engineers and planners and not economists, see Lindsey (2006, p. 303). Vickrey (1969) distinguished six causes of congestion: simple traffic interactions; multiple traffic interactions; bottlenecks; triggerneck (caused by the queues resulting from a bottleneck causing delay to vehicles not intending to use the bottleneck facility); network and control congestion (resulting from a lack of traffic control measures) and general density (this may well overlap with the other causes of congestion). What is clear in Vickrey's work is that he considers there to be a relation between traffic speed and flow (measured in vehicles per unit of time passing a given point). In his work, the objective of policy should be to charge vehicles the social marginal cost of vehicle use. It is interesting that one of Vickrey's (2008) last statements on road congestion could be interpreted as density rather than flow as being the cause of congestion and hypercongestion.

### 2.4 The Standard Model of Road Congestion

Walters (1961) provided a theoretical and empirical study of the marginal external congestion cost using the conventional speed flow relationship with a simple assumption of homogeneity of traffic. In this section, we discuss the basic analysis of how the speed flow relationship is used to estimate the marginal external congestion cost. Since then, many transport economists have referred to the curve

as the conventional speed flow relationship. This conventional static model is examined following the presentation given in Verhoef (2005).

In terms of understanding the conventional speed flow relationship, it is useful to follow Haight (1963) and make use of the basic definition of road traffic flow as the product of speed and density

$$F = V\ D \tag{2.4.1}$$

where $D$ - Traffic density, number of vehicles on a given road

$V$ - Traffic speed, kilometres per hour

$F$ - Traffic Flow, vehicles-kilometres per hour

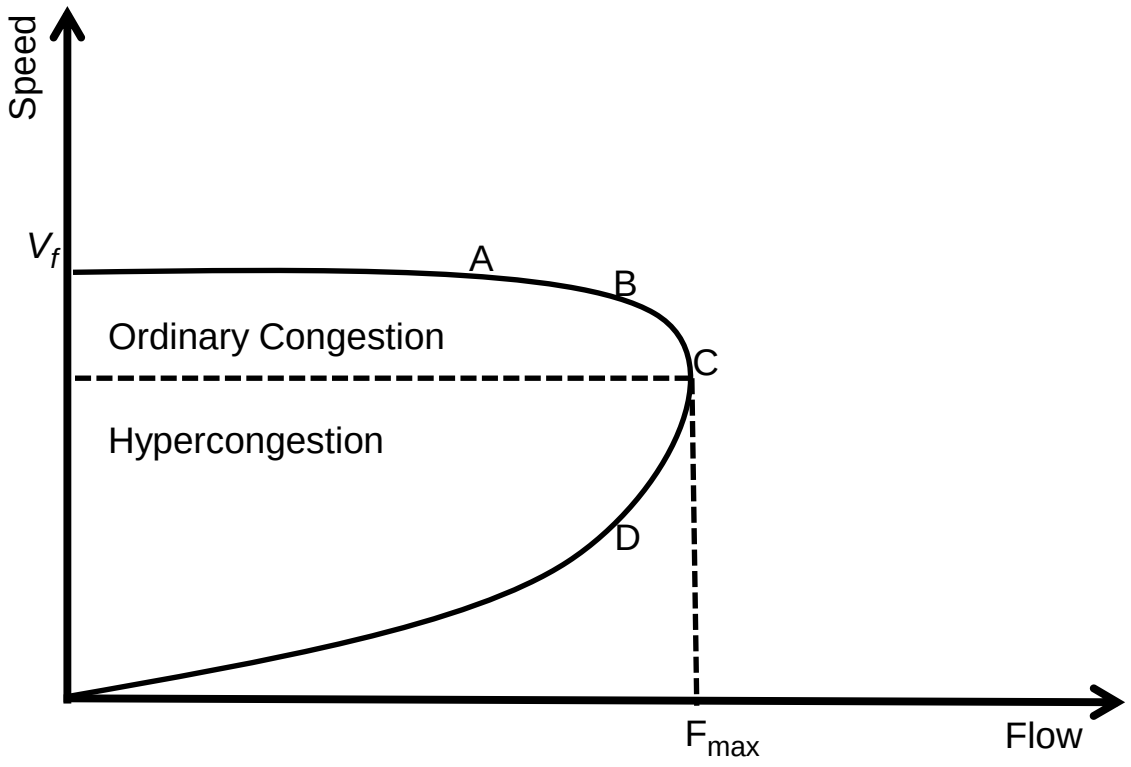


**Figure 2.4.1 The Conventional Speed flow Relationship**

The Figure 2.4.1 gives the commonly drawn standard speed flow relationship (see Walters, 1961; Johnson, 1964; Newbery, 1987, 1989 and 1990; Button, 1993; Verhoef, 1999; Lindsey and Verhoef, 2000b; Verhoef, 2001b; Small and Chu, 2003; Quinet and Vickerman, 2004; Lo and Szeto, 2005 and Small and Verhoelf, 2007, for example) The on its side U shaped curve was originally suggested by Greenshields (1935). In reality, the propagation of speed flow relationship is not as simple as implied by the diagram, but more complex. The shape and location of the curve is determined by a number of factors, i.e. the number and width of traffic lanes, grade, road curvature and slope, weather conditions, driving habits and vehicle types (Small and Keeler, 1977, Button, 1993 and Lindsey and Verhoef, 2000). In addition, Small and Verhoef (2007, p. 72) point out that "later studies revealed that this parabolic shape is less accurate for larger highways, where instead speeds often remain constant, or nearly so, over a substantial range of flow levels." i.e. between $V_f$ and A in Figure 2.4.1.

The conventional speed flow relation consists of two distinct parts. The upper negatively sloped curve represents ordinary congestion and the lower curve with a positive slope represents hypercongestion (Lindsey and Vehoef, 2000a and Small and Verhoef, 2007). In this section, only ordinary congestion is examined with the hypercongestion being considered in a later section.

Let us look at the top part of the diagram in Figure 2.4.1. For the upper part of the curve and over the initial flow range, vehicles travel at a free-flow speed $V_f$ and remain at this speed throughout the range as flow increases. As soon as the slope becomes negative, congestion occurs. However, as long as the effect of an additional vehicle outweighs the reduction in speed, the flow still increases. For

instance, in Figure 2.4.1, when flow increases from *A* to *B,* speed falls but flow increases. Throughout the top part of the curve as flow increases, the underlying cause is the greater vehicle density. This relationship between speed, flow and density lies at the heart of some of the debates on the correct measure of the marginal external congestion cost. When the effect of an additional vehicle exactly balances the decrease of speed, the flow reaches its maximum level at the road capacity.

The conventional economic analysis employs the engineering concept of speed flow relationships for estimating marginal external congestion cost with a simple assumption of traffic homogeneity. But traffic characteristics in reality are more complicated (Walters, 1961, and Verhoef, 1999). Walters (1961) defends the simple assumption of homogeneity as useful for the theoretical development and empirical analysis. Hence, the basic analysis assumes uniform traffic travelling on a constant capacity road. In addition, all road users are assumed homogeneous, i.e. identical vehicle types and driving behaviour.

The basic speed flow model has been used by a large number of studies to derive and estimate the marginal external congestion cost (MECC). The approach is explained using the example of Newbery (1989). A simple approximation of the average travel cost per kilometre of a vehicle is

$$AC = a + \frac{b}{V} \qquad (2.4.2)$$

Where $b/V$ - the time cost (thus, b is value of time) ,

$a$ - the resource cost.

The economic literature on the parameter *b*, the value of time, is considered in a later section.

The total cost of a flow of *F* vehicles per hour is

$$TC = AC\ F \tag{2.4.3}$$

If an additional vehicle-kilometre is added to the flow, the total social cost is increased by

$$\frac{dTC}{dF} = \frac{d\left(AC\ F\right)}{dF} = AC + F\ \frac{dAC}{dF} \tag{2.4.4}$$

From (2.4.4), the first term is the private cost borne by each vehicle and the second is the *MECC* borne by other vehicles.

From (2.4.4), the *MECC* is given by

$$MECC = \frac{b}{V}\left(\frac{-F}{V}\ \frac{dV}{dF}\right) \tag{2.4.5}$$

Newbery (1989, p. 173) gives a simple explanation of the *MECC* in (2.4.5) as “a measure of by how much an extra vehicle-kilometre increases the total time costs of other travellers making their total trips, once they have had time to adjust their travel patterns to the extra traffic”. In terms of the new model developed in Chapter 3, it is notable, see Newberry (1989, p. 173), that the *MECC* is the product of the average cost and the elasticity of flow with respect to speed.

Despite the conventional analysis explained above, there are debates whether flow or density should be the correct basis for estimating *MECC*. Else (1981 and 1982), Evans (1992a and 1992b), Hills (1993) and Ohta (2001a and 2001b) argue

that some form of either density or trips not flow should be the basis of measuring traffic demand. However, Nash (1982) and Verhoef (1999 and 2001a) support the conventional idea of using traffic flow.

Referring to the identity in (2.4.1), Else (1981, p. 220-221) explains that “a decision by an individual to use a road is essentially a decision to add to the number of vehicles on that road. That addition certainly adds to the traffic density, but whether or not it increases the traffic flow depends on the volume of traffic already on the road.” As noted by Nash (1982), this proposition leads Else to three important conclusions, one of which relates to hypercongestion and the debate on this issue is considered in a later section. Firstly, an addition to the number of vehicles on the road causes external cost to later traffic and this is difficult to estimate. Nash (1982) agrees with the latter comment. However, the issue of whether longer journeys times for later traffic are strictly an external cost is complex. For traffic on the road segment at the same time as the additional vehicle, the reduction in speed represents an external cost as such vehicles have to spend longer on the road at some other time to complete their journeys. This in turn creates an externality that is caused by the increased density of these vehicles at other times than the additional vehicle is on the road. This argument is pursued in full in Chapter 3 and Chapter 4. Carey and Else (1985) apply Else’s earlier analysis (1981) to the case of a road with multiple segments. However, the issue identified about what is an externality and what is not is not analysed. Secondly, “the marginal [social] cost of increasing the traffic flow …. must be greater than the marginal [social] cost of increasing the number of vehicles since a given increase in traffic flow requires a proportionately greater increase [than] in the number of vehicles” (Else, 1981, p. 222). Thus, he argues that conventional

estimates of the marginal external congestion costs given by (2.4.5) are over estimates. Whether, this comment is correct and the extent of the overestimate is investigated in Chapter 3 and Chapter 4. Nash (1982, p. 297) has argued that Else is wrong and that “demand curves in economics are conventionally measured in terms of a desired flow (that is, quantity per unit of time)”. Else (1982, p. 300) replied that “under congested conditions, an increase in demand (e.g. as represented by the entry of one additional vehicles to the road) leads to a less than proportionate increase in the flow of output and a consequent spreading out of its effect over time”. This debate was not clearly resolved. It is helpful to note that, in congested conditions in the top part of Figure 2.4.1, within a given period of time, an increase in flow is achieved with an increase in density, a fall in speed and a reduction in the distance travelled by each vehicle. Thus, for a vehicle-kilometre increase in total distance travelled, the additional vehicle has to travel more than one kilometre and the existing vehicles travel a little less. Consequently, estimating the *MECC* using an increase in flow requires the additional vehicle to travel more than one kilometre in this period. This might be taken as implying a lower external cost than the conventional analysis. However, the other vehicles have to make up the reduced distances travelled in this period by travelling in other periods. This additional travel represents an additional external cost and is shown in Chapter 3 ( assuming constant speed and density) to give a *MECC* that is the same as the conventional analysis.

It is also relevant to consider the decisions made by road users that determine road congestion. Road users have preferences over making completed journeys at particular times. Their choices are constrained by the average speed at different times of the day. The user decides upon at which time to make a journey or not on

the basis of these preferences and constraints. Thus, density or flow is not strictly demand but is derived from these choices. Travel takes place across different time periods with different speeds. The consequence of this is that it is not possible to represent congestion using a simple single two dimensional diagram for one period. This last point explains some of the unresolved debates on whether the correct measure of demand is flow or the occupation of the road at one point of time.

Evans (1992) argues in support of Else's view using a diagrammatic approach. The major purpose of Evans is to show the flaws in the conclusions of De Meza and Gould (1987) that monopoly ownership of a common property resource may make all users better-off and that an appropriate toll may make all users better-off (even if they gain nothing from the toll revenue). Perhaps surprisingly, Evans provides no references to the Else (1981 and 1982) and Nash (1982) debate though many of the points at issue are similar.

Evans (1992) makes a welfare economic analysis of road pricing and considers the importance of congestion functions. The analysis led Hills (1993) to criticise the one period static model that Evans used. In one period, there is no room for additional time delays caused by congestion to take place and movement from off-peak to peak demand conditions is impossible. Hills (1993, p. 96) argues that demand is for "desired trips" and "the time taken to reach destination is simply as long as it takes". Thus, Hills (1993, p. 97) suggests that the analysis of road congestion should move "from a *flow-based* measure to a *trip-based* measure". Evans (1993) response was to agree with the inability of a one period model to capture changes in demand over time. However, Evans (1993, p. 102) criticised Hills use of the absolute measure of demand "trips accomplished" as a rate

measure is required. Evans suggests that a vehicle kilometre measure of demand is better than a vehicle trip measure as the former allows for homogeneity and aggregation.

Finally in this section on speed flow models, it is appropriate to consider the use of such models in government transport planning decision making. Taylor, Bourne, Nottley and Skrobanski (2008, p.1) note that “while empirical models suitable for economic appraisal were formulated for the UK’s COBA .... and the US Highway Capacity Manual, in recent years. the ‘fundamental’ relationship between flow, speed and density has been questioned”. However, in the UK, US and many other countries, the speed flow relationship lies behind much transport planning and appraisal. In the UK, the speed flow relationships used in the COBA (COst-Benefit Analysis) are estimated multi-linear functions used for calculating time savings of proposed transport schemes, see Department of Transport (2002). They are average speeds in steady state conditions. Different road types are considered from small rural carriageways to motorways and different road conditions are allowed for, e.g. hills and curves. They do not consider hypercongestion explicitly as somewhat arbitrary minimum speeds are assumed. When the COBA procedure identifies that a segment of a road network is at “overcapacity”, the COBA programme produces an overcapacity report and a practical approach is adopted to making decisions about the need for investment in this part of the network. This latter approach appears to make no direct use of speed flow relationships. The UK Department of Transport’s WEBTAG (2013) pages are their guidance and information on transport modelling and appraisal and makes extensive use of speed flow relations through the use of the COBA procedure. A precursor and

more local use of speed flow relationships can be found in the Midlands Regional Transportation Model, see Wootton (2004).

The various editions of the US Highway Capacity Manual (1950, 1965, 1980, 2000, and 2010) represent a continuing attempt to improve the modelling of traffic flows and speeds on US freeways, and urban and rural roads. These improvements are still based on speed flow relationships for all types of roads and different characteristics, e.g. slopes and curves. However, a departure in the later r editions is the use of dynamic traffic modelling, an emphasis on modelling congestion, consideration of intersections, road corridors and connections. However, there is no explicit consideration of hypercongestion in the reported speed flow relationships and curves.

## 2.5 Empirical Analysis

In practice, the conventional negative speed flow relationship is a common tool for estimating the marginal external congestion cost. The empirical studies of speed flow analyses can be classified into two types: the standard speed flow relationship and the network traffic simulation model. The studies of the standard model directly estimate the speed flow relationship on an assumed homogeneous road segment. Whereas, the network simulation models employ different speed flow relationships on different parts of the road network to obtain estimates of the overall congestion.

In this section, we provide a non-exhaustive list of empirical studies of economic analysis of marginal external congestion time cost (METC) using the conventional speed flow relationship in Table 2.5.1.

The studies examined the marginal external congestion time costs on highway and urban streets of various countries across the world (i.e. the USA, the UK, Canada, Belgium, Singapore and India). The estimated highways marginal external congestion time costs vary between 0.0002 and 0.19 hour/vehicle-kilometre at speeds of 64-170 kilometre/hour. Whereas the estimated urban marginal external congestion costs vary between 0.0005 and 1.2 hour/vehicle-kilometre at speeds of 8.5-72 kilometre/hour. Obviously, the results show that the economic cost of congestion in urban areas is more serious than on highways and that there is a great deal of variation in the highway and urban marginal external congestion time costs.

Data underlying the investigation of some speed flow relations are given in Figures 2.5.1 to 2.5.7 for both highways and urban roads. Again, it is notable that there is a large degree of variation in the data with the fitted curves rarely being a good fit. In these Figures and fitted relationships, and the earlier reported empirical investigations, it is of note that there is little or no diagnostic testing of the estimated relationships in terms of testing for heteroscedasicty, normality of errors and functional form. Such diagnostic testing is standard in modern microeconometrics, see for example Greene (2008) and Wooldridge (2009).

## Table 2.5.1 Empirical Studies of Economic Analysis of Congestion Using the Conventional Speed Flow Relationship

| No. | Author | Data | Purposes | Ranges of Estimated METCs (hour/vkm) |
|---|---|---|---|---|
| 1. | Walters (1961) | Expressways in Charleston West Virginia, the USA, 1944 | Estimating speed flow relationship to give the marginal external congestion time cost. | 0.013-0.167 for peak periods (speeds: 34-60 km/h ) |
| 2. | Johnson (1964) | Arterial streets in Pennsylvania cities, the USA, 1961 | Estimating speed flow relationship to give the marginal external congestion time cost. | 0.004-0.172 for peak periods (speeds: 24-60 km/h) |
| 3 | Smeed (1968) | Streets in Centre of eight towns in the UK including the central of London, 1966 | Estimating area-wide model of speed flow to give the marginal external congestion time cost. | 0.000-1.181 for peak periods (speeds: 8.5-39 km/h) |
| 4 | Small and Keeler (1977) | Radial commutation expressways in San Francisco Bay Area, the USA, 1968 | Estimating speed flow relationship to give the marginal external congestion time cost. | 0.003-0.045 for peak periods (speeds: 80-95 km/h) |
| 5 | Boardman and Lave (1977) | Interstate highways in Maryland and Massachusetts, the USA, 1970 | Estimating speed flow relationship to give the marginal external congestion time cost. | 0.0002-0.036 for all day periods (speeds: 85-170 km/h) |
| 6 | Inman (1978) | A major four-lane Washington D.C. commuter artery, the USA, 1968 | Developing generalized congestion functions by using speed flow relation. | Using Drew's function :0.001-0.187 for peak periods (speeds: 65-127 km/h)<br><br>Using Boardman and Lave's function:0.003-0.070 for peak periods (speeds: 65-127 km/h) |
| 7 | Dewees (1979) | Road network in Toronto, Canada, 1978 | Developing a traffic simulation model using a speed flow relationship to produce new estimates of congestion costs on specific streets. | 0.000-0.680 for peak periods, inbound traffic (speeds: 12-38 km/h)<br><br>0.0001-0.095 for peak periods, outbound traffic (speeds: 28-47 km/h) |
| 8 | Newbery (1989) | 13 towns and cities in the UK, 1978 | Estimating speed flow relationship to give the marginal external congestion time cost. | 0.02-0.04 for peak periods (speeds: 21-33 km/h)<br><br>0.01-0.03 for off-peak periods (speeds: 24-38 km/h) |
| 9 | Newbery (1990) | Urban streets in Great Britain, 1985 | Estimating speed flow relationship to give the marginal external congestion time cost. | 0.023-0.054 for peak periods (speeds: unspecified)<br><br>0.013-0.043 for off-peak periods (speeds: unspecified) |
| 10 | Mayeres et al. (1996) | Brussels in Belgium, 1993 | Estimating speed flow relationship to give the marginal external congestion time cost. | 0.0005-0.1800 for peak periods (average speed: 38.2 km/h) |
| 11 | De Borger et al. (1996) | 6 cities in Belgium, 1989 | Estimating speed flow relationship to give the marginal external congestion time cost. | 0.0357 for peak periods (average speed: 10 km/h)<br><br>0.0024 for off-peak periods (average speed: 30 km/h) |

| | | | | |
|---|---|---|---|---|
| 12 | Sansom et al. (2001) | Varieties of Road types in Great Britain, 1996 | Estimating speed flow relationships to give the marginal external congestion time cost | 0.0160 for all day periods in Motorways<br><br>0.1077-0.1078 for peak periods on central of major urban areas<br><br>0.0287-0.0304 for peak periods on non-central of major urban areas<br><br>0.0057-0.0105 for peak periods on other urban areas<br><br>0.0114-0.0115 for all day periods on rural trunk and principal roads<br><br>0.0016-0.0115 for all day periods on other rural roads<br><br>(Speeds: unspecified) |
| 13 | Li (2002) | Singapore Central Business District Area, 1997 | Estimating speed flow relationship to set different tolls. | 0.0026-0.0038 for peak periods (speeds: 55-72 km/h) |
| 14 | Newbery and Santos (2002) | Urban road networks of 8 towns in the UK, 1998 | Testing the reliability and applicability of linear speed flow relationships in estimating the marginal external congestion cost by using simulation models, SATURN[11]. | 0.007-0.225 for morning peak periods<br>(speeds: 16.4-57 km/h) |
| 15 | The Scottish Executive (2003) | Trunk roads in Scotland, 2003 | Estimating speed flow relationship to give the marginal external congestion time cost. | 0.0013 for peak periods on A-roads<br>(speeds: 64-113 km/h)<br><br>0.0015 for peak periods on M-roads<br>(speeds: 69-111 km/h) |
| 16 | Sen, Tiwari and Upadhyay (2010) | Urban roads in Delhi, India,2005 | Estimating speed flow relationship to give the marginal external congestion time cost. | 0.25 for peak periods<br>(average speed: 24 km/h)<br><br>0.007 for off-peak periods<br>(average speed: 40km/h) |
| 17 | Department for Transport, WEBTAG (2011a) | Road networks in the UK, 2010 | Estimating Marginal External Costs using speed flow relationships provided by the National Transport Model. | 0.047 for A roads<br>(speeds: unspecified)<br><br>0.023 for other roads<br>(speeds: unspecified) |
| 18 | Department for Transport, COBA11 (2012a) | 13 towns in the UK, 1976 | Estimating speed flow relationships. | 0.014 for peak periods in central areas<br>(average speed: 21.5 km/h)<br><br>0.008 for peak periods in non-central areas<br>(average speed: 36.5 km/h) |

Notes: (1) vkm stands for vehicle-kilometre. (2) km/h stands for kilometres per hour.

[11] SATURN represents Simulation and Assignment of Traffic to Urban Road Networks, a software package developed by Institute for Transport Studies at Leeds University.

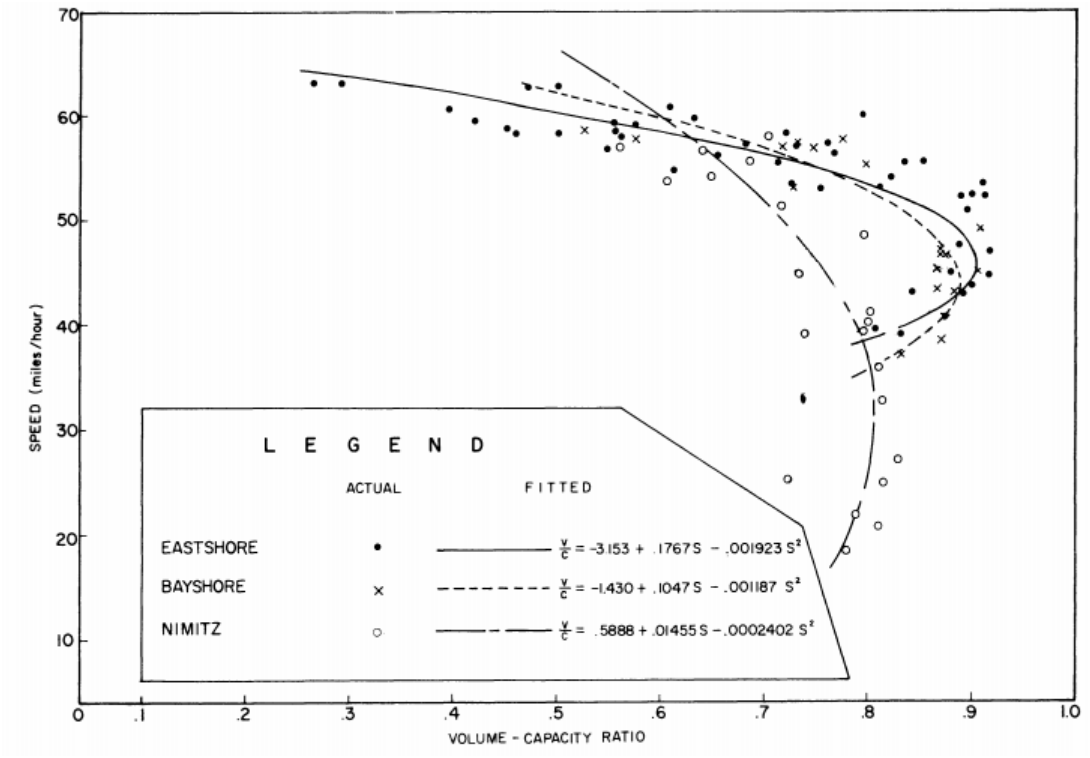


**Figure 2.5.1 Radial Commutation Expressways in San Francisco Bay Area, USA, 1968 [in Small and Keeler, 1977: p. 12]**

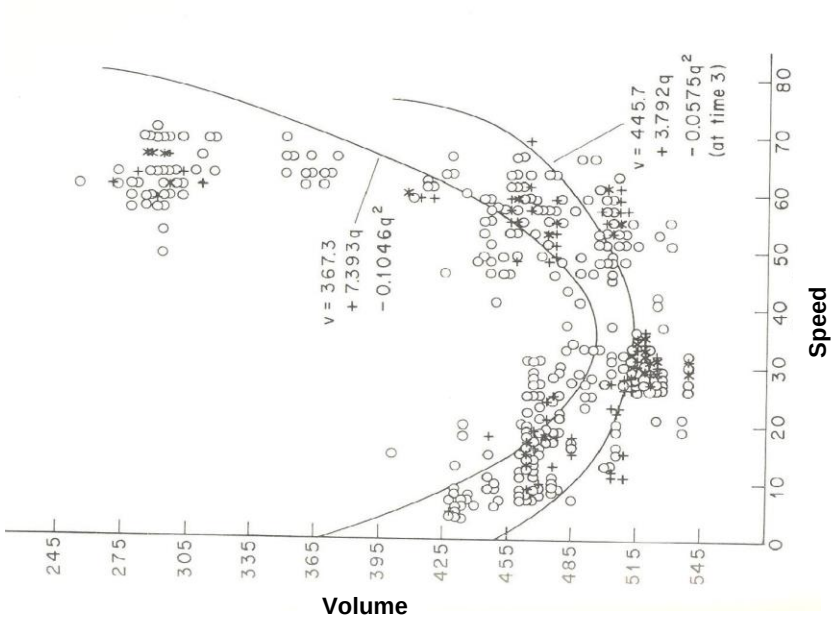


**Figure 2.5.2 Interstate Highways in Maryland and Massachusetts, USA, 1970 [in Boardman and Lave, 1977: p. 346]**

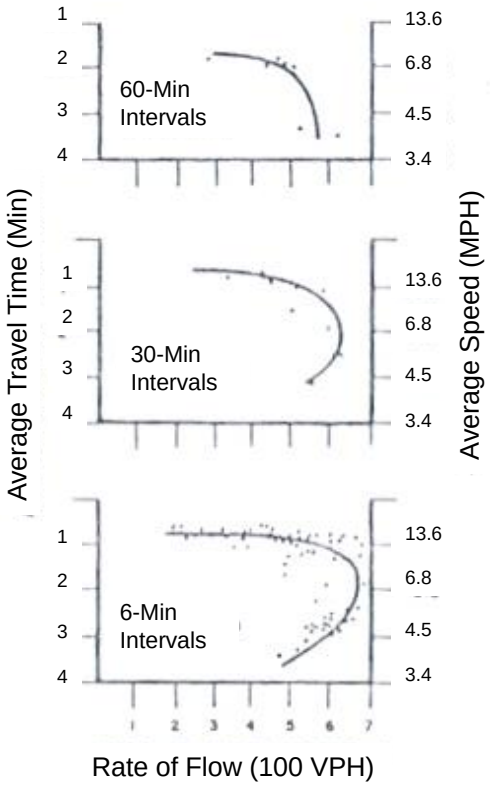


**Figure 2.5.3 Signalized Streets in Intermediate Urban West Charleston, USA [in HCM 1965: p. 67]**

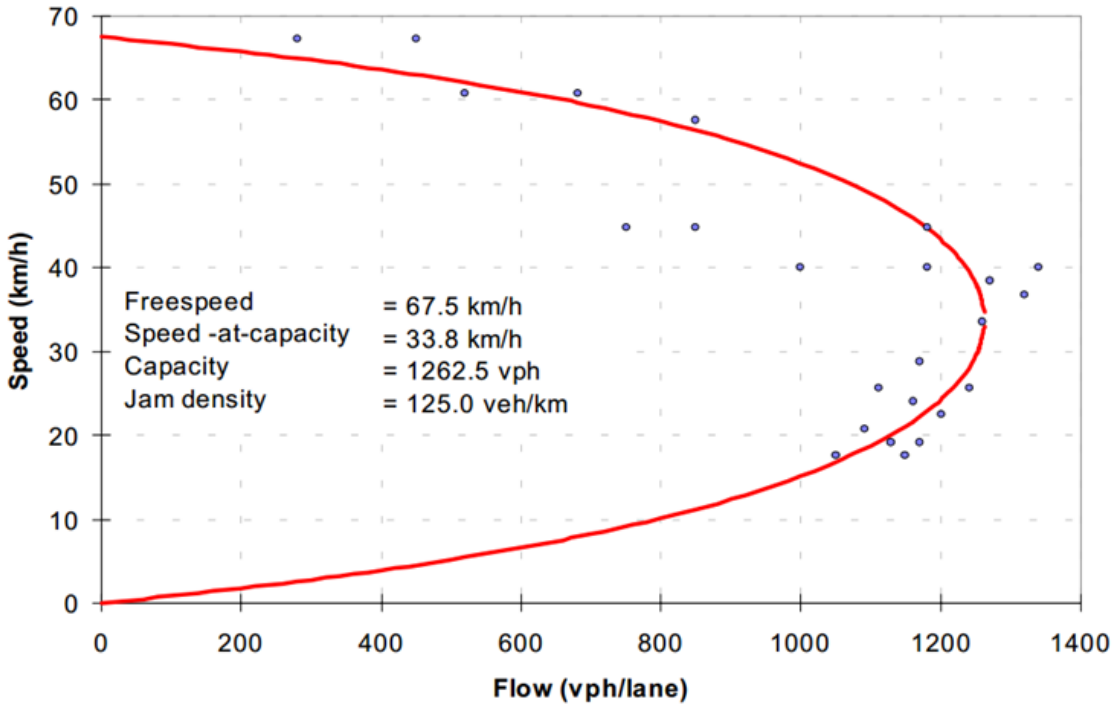


**Figure 2.5.4 The Caldecott Tunnel in California, USA**
**[in Aerde and Rakha 1995: p. 6]**

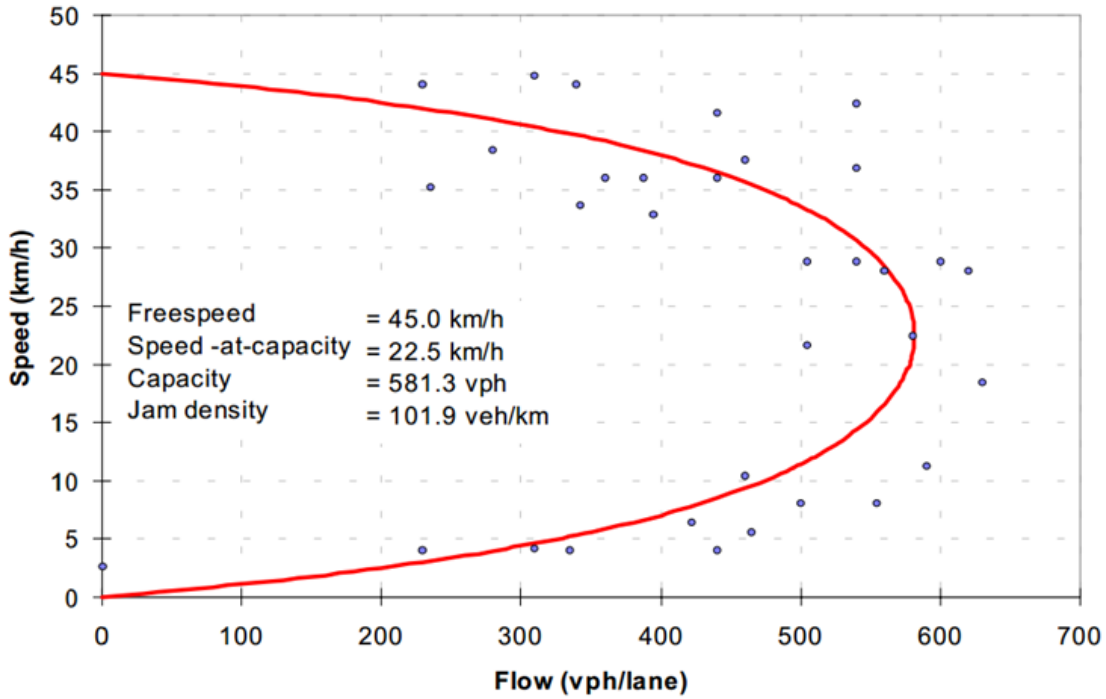


**Figure 2.5.5 An Arterial Street in California, USA**
**[in Aerde and Rakha, 1995: p. 6]**

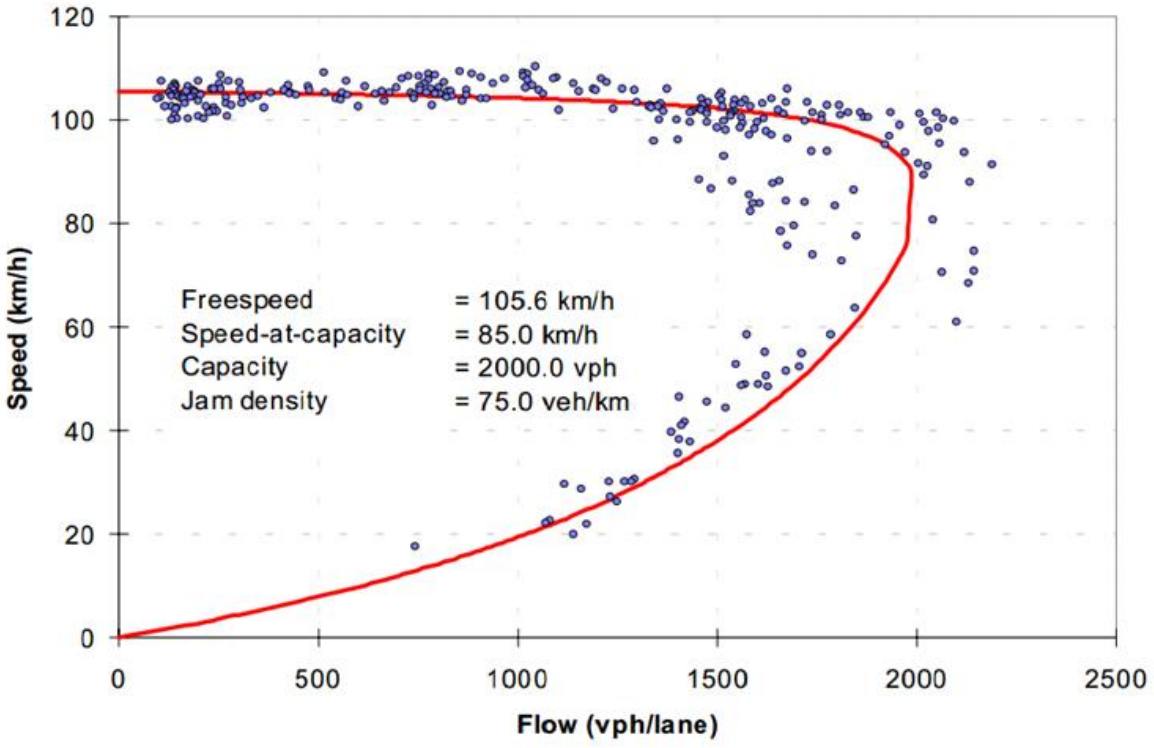


**Figure 2.5.6 A Freeway in Amsterdam, Netherlands**
**[in Aerde and Rakha, 1995: p. 7]**

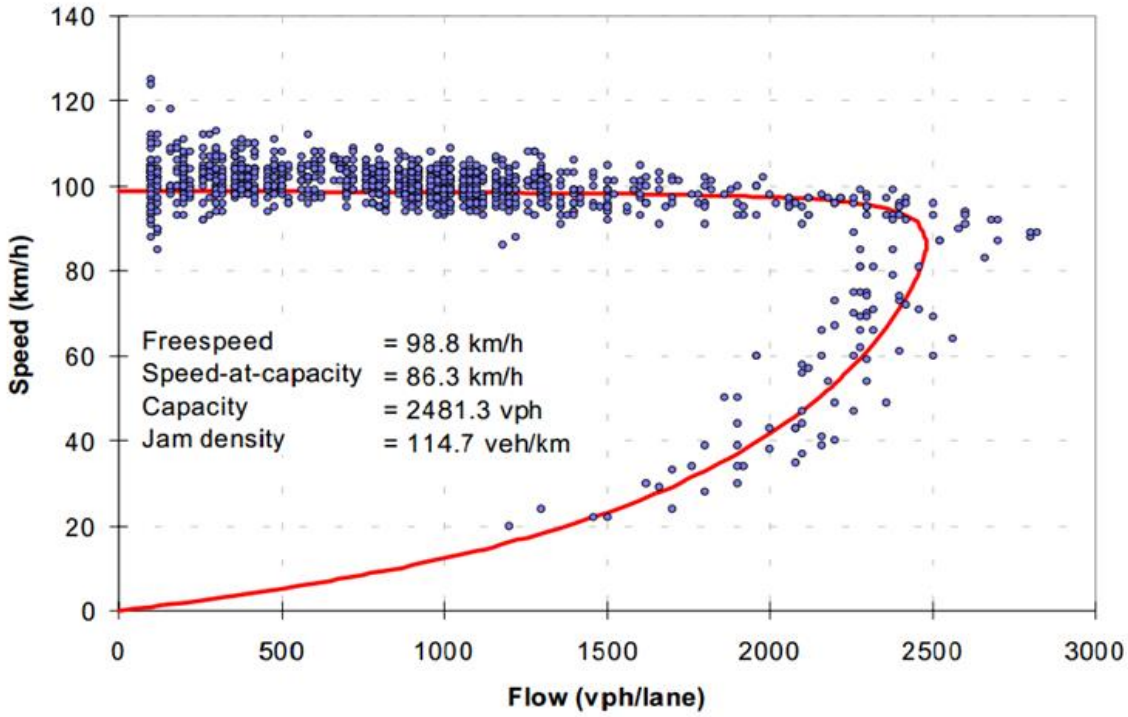


**Figure 2.5.7 A Highway in Toronto, Canada**
**[in Aerde and Rakha 1995: p. 7]**

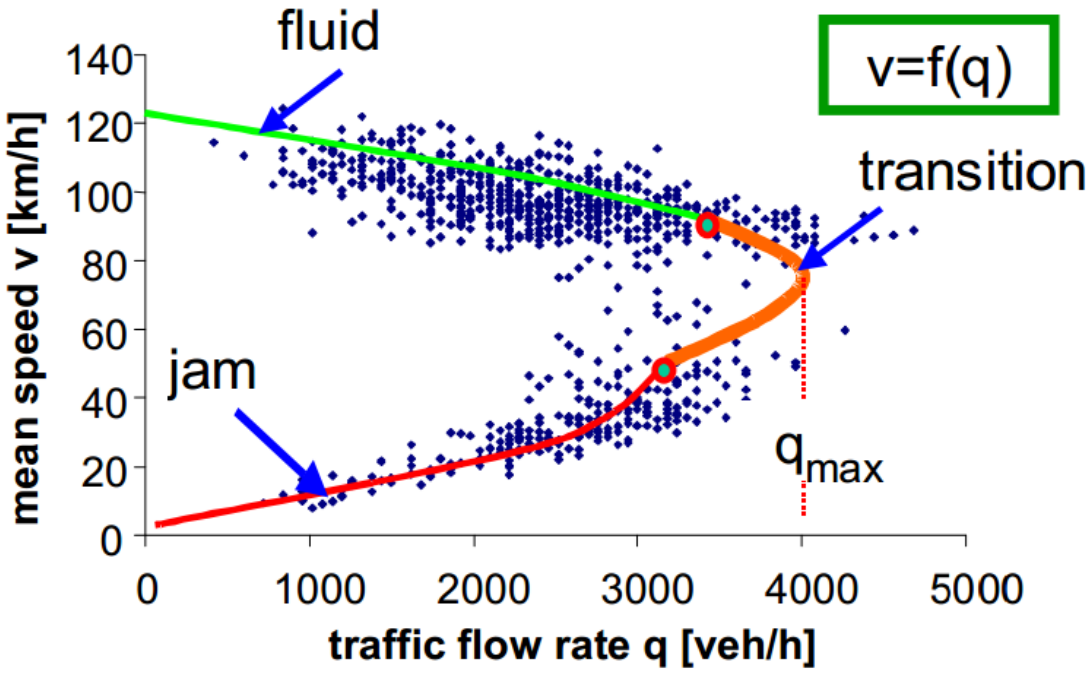


**Figure 2.5.8 Two Major Freeway Interchanges in Germany [in Wu, 2002: p. 869]**

## 2.6 Dynamic Models

The static one period model considered above has been a tool used to estimate the external costs of congestion. However, it is clear that static models fail to capture three important characteristics of road use related to congestion, see De Palma and Fosgerau (2011). Firstly, congestion varies across the time of day, week and season. It is important to model how traffic moves between different patterns of flow. If a charge for the use of roads is made, this charge should vary with the level of congestion and this can only be modelled in a multiperiod model. Secondly, road users have preferences for departure and arrival times that are in addition to the time costs of making a journey. These preferences can only be

captured in a multiperiod model. Thirdly, deviations in arrival times caused by unexpected traffic levels can only be modelled within multiperiod models that incorporate demand uncertainty. This latter issue is important but not pursued in this study.

Thus, in order to develop a theoretically sound model of road congestion, it is important to consider multiperiod models of traffic flow, density and speed. These types of models are often called dynamic though the exact meaning of the term dynamic model may vary between different authors.

Vickrey (1969) was the first economist who considered a dynamic model based upon the idea of a bottleneck. This model was taken up and extended by Arnott et al. (1993). The basic dynamic model consists of one route between two points with one bottleneck. All users wish to arrive at these destinations at the same time and arriving early or late incurs a cost to users. The bottleneck constrains users to a maximum flow at this point. Trip scheduling time and delay costs are endogeneous variables in this model. With a flow in excess of the bottleneck maximum flow, queues arise and are waste of the resource time.

This Vickrey model (1969) can be extended to incorporate optimal and second best tolls (Arnott et al., 1993a, and Arnott et al., 1990), financing (Verhoef and Mohring, 2009), inelastic and elastic demand (Arnott et al., 1993a), scheduling preferences ( Vickrey, 1973, and Tseng and Verhoef, 2008), variability in journey times (Fosgerau and Karlström, 2010), parking costs (Arnott et al., 1991), and networks (Arnott et al., 1993b and Merchant and Nemhauser, 1978).

In the context of dynamic models, Verhoef (1999, p. 343) brings up the issue of the transition between different phases of demand and admits that it is complicated to incorporate the transition phase to the analysis. Indeed, he is very aware of the unrealistic assumption that must be used and explains that “road traffic congestion in reality is a complicated dynamic process, and the analyst studying congestion and congestion pricing is soon confronted between using either an “as realistic as possible” modeling approach, in which analytical solutions are often difficult to obtain, or to apply a simpler representation of reality, allowing analytical solutions and the derivation of more or less general insights into the economic principles behind the problems studied.”

The motivation of the present thesis is to derive a model of the economics of congestion that allows the external costs of congestion to be estimated. Whilst there exist a number of stylized analytical dynamic models that provide insights into the congestion process, actual estimation of the congestion costs requires numerical simulations that rely heavily on the underlying assumptions and assumed parameter values. Thus, the next section attempts to provide a dynamic model that can be used to estimate empirically the costs of congestion and explain hypercongestion.

The literature review shows that the past estimation of the congestion costs is often based upon a one period model with speed being implicitly determined by flow. It is recognised in the transport economics literature that dynamic models of road traffic are much more realistic and allow the consideration of many important characteristics of road transport. Many of the dynamic models consider the relation between speed and flow. Additionally, there is not a simple analytical manner in

which to estimate congestion costs from dynamic models. It is necessary to rely on computational simulations to estimate congestion costs (see Merchant and Nemhauser, 1978; Benakiva et al., 1986; Arnott et al., 1993b; Verhoef 2001b; Kuwahara, 2007; De Palma and Fosgerau, 2011, for example).

## 2.7 The Density Based Congestion Model

Thus, it is important to develop and construct models of dynamic traffic behaviour that allow the derivation or at least consideration of explicit functional forms that represent the costs of congestion. The basics of such a model are constructed here and used for two purposes. Firstly, it is used to examine and understand the economics of hypercongestion in a later section of this Chapter. Secondly, in later chapters, it is shown how the dynamic model constructed here can be estimated and used to give functional forms and simulations for the costs of ordinary congestion, hypercongestion and congestion impacts of different types of vehicles.

A potential road user can make many different journeys at different times of the day. The user is assumed to have perfect knowledge of the densities of traffic by other users on all roads at all possible times. Thus, the individual user chooses journeys on the basis of her preferences for journeys and their timing, and the costs of journeys. In this sense, the model is very similar to a standard microeconomic model for any good or service. The difference is that the user is choosing occupation of the road in different time periods to ensure completion of the chosen journeys. Thus, the desired service is occupation of the road and these demands are highly complementary as the user occupies the road in adjacent time periods until the journey is completed.

Each chosen journey is made across a number of time periods. We take the length of these time periods as infinitely small and the integral of the speed across time is just equal to the length of a journey. Speed is determined by the time and location on the road. These circumstances are partly depicted in Figure 2.7.1.

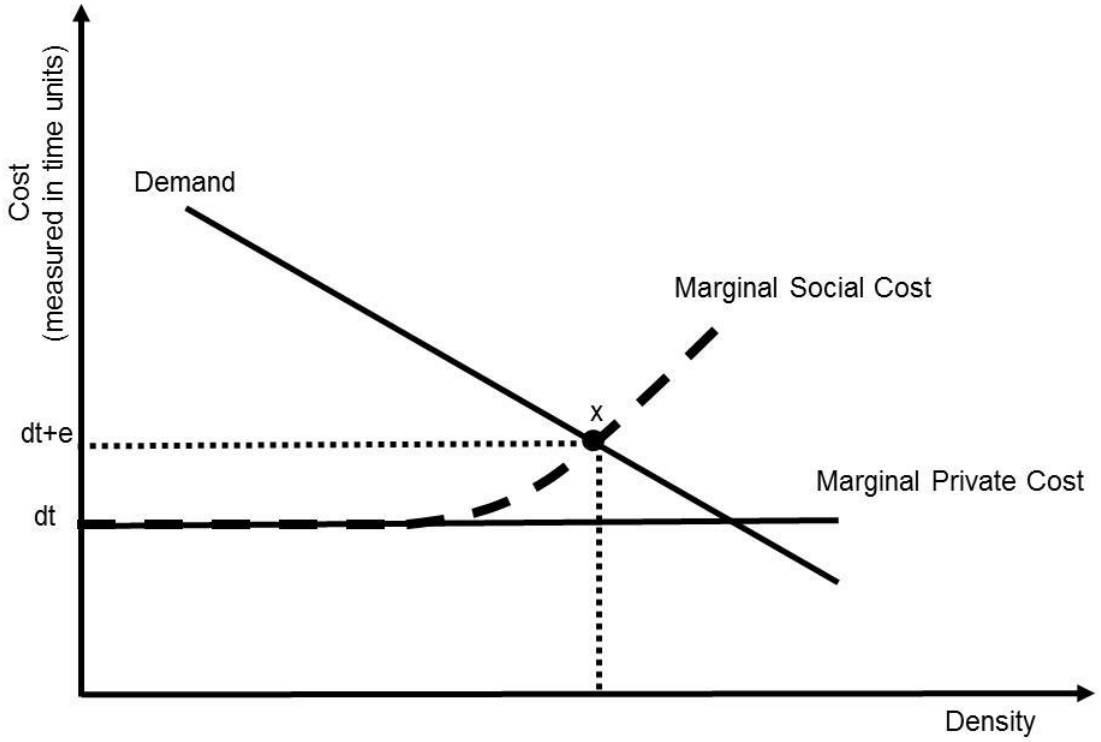


**Figure 2.7.1 One Period of the Density Based Congestion Model**

Thus, for each period and location, we have demand schedule for road occupation. The cost of the journey in each time period is fixed at the length of the time period (dt), an implicit assumption in the speed flow analysis of congestion but rarely made clearly. The external effect of congestion is modelled by an additional road user causing other users to have to spend longer on the road with

either earlier start times or later arrival times.[12] Thus, the external effect is that other users have to spend longer over the journey in different time periods. This is the great advantage of this dynamic model over the speed flow one period model. In the latter model, there is implicitly no other time periods in which other users can travel and nowhere for the time congestion cost to exist.

The extent of this external effect depends on the speeds in the periods in which the other users make up for the lower speed following from the additional user occupying the road in the period under consideration. This information will not in general be known but can be approximated as is shown in Chapter 3.

The new model emphasises that there are two external effects from congestion. The first is the longer time taken over journeys. The second is that the other users change their timing of making journeys or may not make the journey at all. With an additional road user, every other user has to start either earlier or finish later or both or not make the journey. This rescheduling results in less desirable journeys timings and could be analysed in terms of shifts in demand curves. This effect is important but difficult to model and, unfortunately, not pursued directly in the empirical investigations in the following chapters. However, it is important to be aware that this effect is omitted from the speed flow analysis and is important. The effect of deterred demand for journeys is considered but not measured in Chapters 3 and 4.

The model is illustrated in Figure 2.7.1. The flat private average and marginal cost is the time cost of occupying the road and must by definition be fixed. The social marginal cost includes the two external effects of an additional road user

[12] It is possible that the other users may choose different routes or not make the journey. This does not affect the analysis.

occupying the road and making other users journeys take longer (in other time periods) and the costs of displaced and less preferred timing of journeys.

This model is simple and is used to consider past studies of hypercongestion in the next section. More importantly, in succeeding chapters, it is used to develop a dynamic analysis which explicitly models the costs of ordinary congestion, hypercongestion and the congestion impacts of different types of vehicles. The dynamic analysis is then used as the basis for empirical estimation of these congestion costs and the practical limitations of this estimation are examined.

## 2.8 Hypercongestion

Before we review the economic analysis of hypercongestion, it is necessary to make clear what the phenomenon of hypercongestion is. Intuitively, hypercongestion is seen as a situation where too many vehicles are stuck on a road segment. In this situation, the traffic stream flows with a low speed and possibly stops sometimes. In general, hypercongestion is defined here as a situation of a positive speed flow relationship.

In spite of the above discussion of hypercongestion, there are many other views and definitions, some of which are very close to the above definition and some not so close, see for example Small and Keeler (1977), Arnott (1990); Button (1993); Lindsey and Verhoef (2000b); Small and Chu (2003); Verhoef (2000 and 2003); Quinet and Vickerman (2004) and Lo and Szeto (2005).

What causes hypercongestion is not completely clear. Arnott (1990, p. 200) explains that “hypercongestion occurs as a transient response of a non-linear

system to a demand spike." In response to Arnott, Small and Chu (2003, p. 323) add that, "hypercongestion occurs when a capacity limit is exceeded somewhere in the system. As a result local queuing begins, which becomes more severe the more cars are added to the input flows. Queuing adds to trip times beyond what is portrayed by the instantaneous speed flow relationship. Of course, this condition cannot persist indefinitely, for travel time would rise without limit. Demand must at some point fall back below the level that caused queuing in the first place." Likewise, Lindsey and Verhoef (2000b) explain hypercongestion with in terms of queuing as this commonly occurs in a state of hypercongestion. Relevantly, Verhoef (2003, p. 540) simply describes that "hypercongestion arises as soon as the capacity is not constant." However, Verhoef (2003, p. 533) considers that "a hypercongested equilibrium in a stationary state model is one in which the traffic density is so high, and consequently the speed is so low, that the traffic flow is below the maximum possible flow for the road (its capacity), and where the speed is below the maximum possible speed at that traffic flow." Similarly, Button (1993, p. 111) and Lo and Szeto (2005, p. 706) indicate hypercongested flows are lower than the maximum level. In addition, Lindsey and Verhoef (2000b) suggest that hypercongestion occurs with low flows and low speeds, a view very close to the definition used here. Additionally, Small and Chu (2003, p. 320) indicate that "hypercongestion is a positive speed flow relationship on the conventional diagram." Quinet and Vickerman (2004, p. 125) consider the positive speed flow relationship as an unstable flow with low speeds. They explain that this circumstance arises with a lot of stop and go conditions at high traffic densities. This is similar to Small and Keeler (1977, p. 11), who suggest that " the backward-

bending portion is a result of stop-and-start driving at bottlenecks during congested periods".

Now let us consider the lower speed flow relationship in Figure 2.4.1 again. Before we move on to the detail of the positive speed flow curve, suppose the speed flow gradually relocates from *C* to *D* along the lower curve. This motion demonstrates the situation after the flow reaches its maximum at the road capacity, $F_{MAX.}$ In this situation, an extra vehicle still enters onto the road. As a result, speed is reduced and, as the impact of a decrease in speed outweighs an increase in density, traffic flow decreases simultaneously. This explains how a speed flow moves backward from *C* to *D*. Under these circumstances, if more vehicles keep entering onto the road, the traffic will eventually stand still. This implies a zero speed and flow.

Before we move to the further discussion, let us consider the speed-flow-density relationships corresponding to the identity (2.4.1) in Figure 2.8.1 (this analysis is based upon that of Verhoef, 2005). This diagram exhibits the conventional speed flow on the right panel and the speed-density on the left. We can see that the whole speed-density curve is either flat or negatively slopde and it is a single-valued function. In addition, an increased density is directly associated with flow, either increasing or decreasing. For instance, in the ordinary congestion, where $D_n$ represents a low density, $F_0$ is a traffic flow rate at a high speed $V_n$. Suppose more vehicles keep entering to the road (an increase in density), traffic volume must reach the maximum flow $F_{max}$ at some point in time. After this point, flow starts decreasing and eventually flow falls back to $F_0$ at a low speed $V_h$ (such circumstances represent hypercongestion). Unlike flow, density still increases up to a maximum level $D_h$.

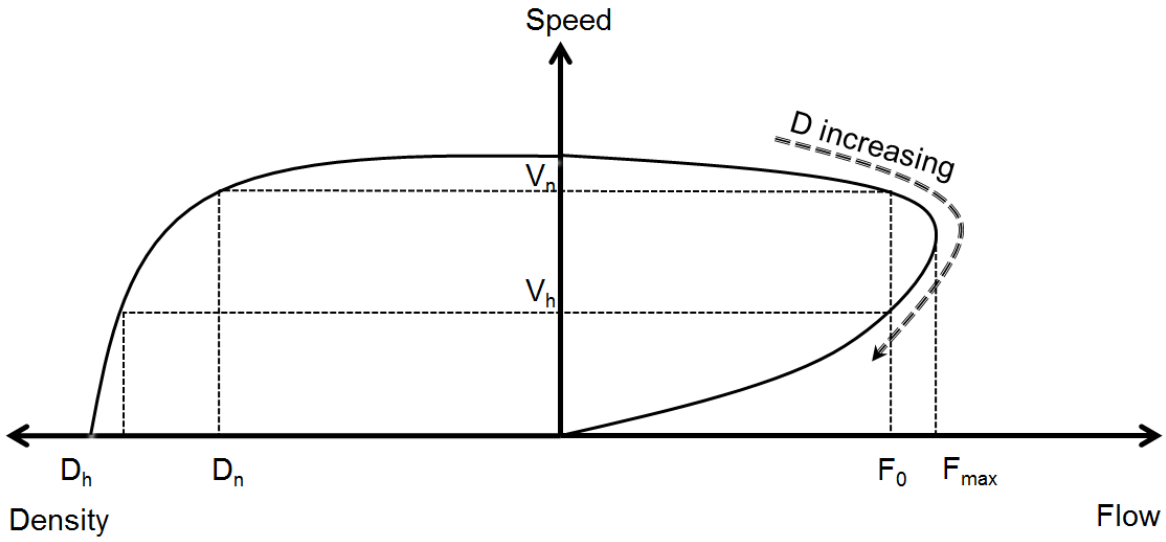


**Figure 2.8.1 Speed-Flow-Density Relationships**

In order to examine the implicit model of Newberry (1987 and 1989), consider the reverse situation on the same positive speed flow curve in Figure 2.4.1. Suppose one vehicle is removed from the hypercongested traffic. With this effect, a speed simultaneously increases with a traffic flow. Correspondingly, the speed flow coordination moves from *D* towards *C*. By considering this conventional diagram, we find Newbery's explanation (1987 and 1989) is not clear. He suggests at the same levels of flow, points *D* and *B*, traffic could immediately switch from a low speed in hypercongestion to a high speed in ordinary congestion. With the above understanding of the conventional analysis, this scenario is not possible.

Although Walters (1961) was the first to consider hypercongestion in a speed flow relationship, the economic analysis of hypercongestion still has not arrived at a consensus (Verhoef, 1999).

The core problem is particularly associated with the positive speed flow relationship. For instance, in Walters (1961), a speed flow relationship is discussed as a time cost function of traffic flow. Correspondingly, Walters (1961, p.

680) comments that “on the backward sloping part of the unit cost curve the marginal social cost is not defined since the change in output is negative”. Else (1981, p. 220) made the argument that “while an equilibrium position may occur on the backward sloping section of the cost-flow, the optimum position cannot, since in that region the change in traffic flow is negative and hence the costs of increasing the traffic flow are effectively infinite.” Another remark by Newbery (1987, p. 18) is “$\frac{du}{dq}$ is positive and the implied externality is negative, which is absurd.” (where *u* denotes traffic speed and *q* denotes traffic flow.) Similarly, Newbery (1989, p. 174) points out that “$\frac{du}{dq}$ is positive on the backward bending part of the curve [and this] implies implausibly negative congestion externalities.”

On the other hand, Newbery (1989, p. 174) identifies that “the backward-bending part of the curve corresponds to dynamic disequilibrium, in which the effect of extra traffic is to cause subsequent vehicles to slow down as the ripple moves back upstream, rather than causing congestion just to neighbouring vehicles.” With a slightly different view than Newbery (1989), Johnson (1964, p. 141) and Small and Chu (2003, p. 220) suggest that the marginal external hypercongestion cost can be an equilibrium, but it cannot be efficient. Furthermore, many economists, for example, Vickrey (1969); Newbery (1989); Arnott Palma and Lindsey (1993); Lindsey and Verhoef (2000) and Small and Chu (2003), point out that the conventional analysis fails in capturing the dynamic characteristics of hypercongestion in terms of place and time. Additionally, Quinet and Vickerman (2004) and Button (1993 and 2004) indicate that the unstable phenomenon in hypercongestion is caused by a lot of stop and go conditions. Likewise, Woensel, Vadaele and Wuyts (2006) refer to a stop and go condition (or a stop-start) as a

non-stationary and a never-stop condition as a stationary state. As a result of this complexity, estimation using the conventional speed flow analysis typically ignores the use of the positive speed flow relationship and only estimates the marginal external congestion cost on the negative slope (Button, 1993; Lindsey and Verhoef, 2000 and Newbery and Santos, 2002).

Ohta (2001a, p. 676) considers the “persistent controversy on the nature of traffic congestion” and argues that flow is a” misleading variable” and that “traffic demand is derived from the need to complete a trip”. The argument is not presented in a clear manner and involves equations, complex diagrams and numerical examples. The important conclusion is that hypercongestion may be an optimal outcome. This argument rests on the view that if demand for completed trips and, thus, density is sufficiently large, then optimal flow could be quite small and hypercongestion is an optimal outcome.

Verhoef’s (2001a) response to the views of Ohta is very strong and in Verhoef (2005, p. 793) he later not surprisingly states that “academia has not yet reached general consensus on the fundamentals that should underlie such analysis [of road traffic congestion]”. Verhoef (2001a) argues strongly that Ohta considers a one period model and ignores the variation in demand across time that must feature in any realistic model of traffic congestion that explains variations between off and on peak periods. Verhoef points out that in periods of hypercongestion stationary state equilibria must be unstable and this makes it impossible to move between hypercongested steady state equilibria. Perhaps most importantly, Verhoef argues that with two possible speeds at each flow, then a hypercongested equilibrium is inferior to the outcome with the same flow but higher speed as the latter has lower costs.

Verhoef's comments led to a response from Ohta (2001b). Amongst other points, Ohta argues that hypercongestion could be an optimal outcome as the same flow with a lower speed would benefit a greater number of drivers. Implicitly, this suggestion is assuming that the occupation of the road by more drivers may convey a greater social welfare. The validity of the argument of hypercongestion equilibrium being efficient is considered below.

The model developed in the previous section is used to examine the Ohta-Verhoef debate. The present study argues that density is a helpful variable with which to analyse congestion. It is shown in the Chapter 4 that hypercongestion occurs when the elasticity of speed with respect to density exceeds one in magnitude – this is the condition for the flow to be below the maximum flow at a low speed, see Figure 2.4.1. This means that if, for example, at levels of density that just reach hypercongestion one could allow each one half of the actual hypercongested density to occupy all the road space by themselves for half the time, all users would travel further in their allotted time. Thus, all users could benefit. This idea can be generalised to show that it is optimal to intervene and restrict travel in hypercongested periods, i.e. when the elasticity of speed with respect to density exceeds one in magnitude, in a manner that can benefit all road users. What is being suggested here is that there potentially exist quota type schemes of use of the road capacity that benefit everyone. This example would seem to contradict an implied externality and efficiency argument in Figure 2.7.1 that the point x is an optimal outcome in spite of it being hypercongested. The effect of an intervention to limit density actually alters the way the road is used to potentially users' benefit. Such quota like policies could be interpreted as raising the demand curve as the journey time cost is reduced and there may also be benefits in terms of more

preferred journey timings. What is clear is that a potential Pareto improvement is possible with a hypercongestion equilibrium and so it cannot be efficient.

The analysis of Chapter 4 that hypercongestion directly implies that the elasticity of speed with respect to density is greater than one in magnitude can be used to show the dynamic instability of hypercongested equilibria. In such an equilibrium, if density increases, the existing road users have to occupy the road for longer in adjacent time periods. The elasticity of speed with respect to density exceeding one (in magnitude) means that the increase in density in adjacent periods must be greater than the initial disturbance in density. Thus, hypercongested equibria are unstable.

It is suggested that though hypercongestion can be an equilibrium, it is very likely to be unstable and is not efficient as other outcomes are superior. Interestingly, it may be the case that policies other than charging could achieve superior outcomes than only first best charging for hypercongestion. These arguments are considered in part in later chapters and it is suggested are worthy of future research. Given the existing academic disputes in this are area and the prevalence of hypercongestion, the suggestions made here are tentative and require further theoretical and empirical analysis.

### 2.9 The Marginal External Congestion Cost of Different Vehicle Types

The conventional economic analysis of traffic congestion generally assumes various homogeneous conditions for traffic. This assumption consists of uniform traffic travelling on a constant capacity road including the traffic stream of uniform

vehicle types and driver behaviour. But traffic in reality has more complicated characteristics (see for example Walters, 1961, and Verhoef, 1999). Traffic stream is actually made up of different vehicle types, e.g. trucks, buses, recreational vehicles and passenger cars. It is important to estimate the effect on congestion of different vehicle types.

In engineering studies, the effects of heterogeneity of different vehicle types are counted and presented in a standardized unit such as passenger car equivalents (PCE) or passenger car units (PCU). Webster and Elecfteriadou (1999, p. 323) and Al-Kaisy, et al. (2002, p. 725) describe PCE as the number of passenger cars displaced in the traffic flow by a truck or a bus, under the prevailing roadway and traffic conditions. Demarchi and Setti (2003, p. 96) explain with a slightly different view that "the PCE of truck is the number of passenger cars that, if replacing that truck in the traffic flow, would have the same effect as the truck on the drivers' perception of the quality of service provided by the facility." Newbery (1989, p. 166) states a similar definition of passenger car units (PCUs) that is a measurement of a congestive effect of a vehicle. He suggests that the measurement is a rough proportion of the vehicle to the road space the vehicle occupies. For example, the congestive effect of a representative car is 1 PCU and a truck is about 2-3 PCUs. Likewise, Santos and Shaffer (2004, p. 165) define that "PCU stands for passenger car unit, a measure for a relative disruption that different vehicle type impose on the network".

In the US, the 1950 first edition of Highway Capacity Manual (HCM) recommends a single factor of 2 to convert the impact of a heavy vehicle as an impact of two passenger cars (Small and Keeler, 1977, and Al-Kaisy et al., 2002). The first published use of the term PCE appears to be in the 1965 edition of HCM, see Al-

Kaisy, Hall and Reisman (2002) and Rahman et al. (2003). Since then, the successive editions of HCM provide sets of PCE of different vehicle types for used as a standard measure in various highway facilities and operations analysis. Thus, HCM, (1994, p. 100) uses the definition of PCE as “the number of passenger cars that are displaced by a single heavy vehicle of a particular type under prevailing roadway, traffic and control conditions”. Similarly, the standard passenger car units (PCU) of different vehicle types used in the UK are provided in the Department of Transport’s Cost Benefit Analysis 11 manual (COBA) for the same purpose. The various factors are light goods vehicle - 1.0, rigid goods vehicle - 1.9, artic goods vehicle - 2.9, and public service vehicle 2.5, Department of Transport (2014). Thus, in the UK, an attempt has been made to differentiate between the congestion impacts of a number of different types of vehicles.

It is suggested here that many but not all studies investigating the effect of different types of vehicle on congestion use numerical simulation to a greater or lesser extent and do not derive their estimate of congestion effect directly from field data. Mallikarjuna and Ramachnadra Rao (2006) argue that passenger car equivalences are too complex to be estimated through analytical techniques and it is common practice top use simulation techniques in estimating values. Shalini and Kumar (2014) suggest that there are eight different methods that can be used to determine passenger car equivalences (PCEs). These methods are briefly considered. The level of service method relates different flows of cars and other vehicles to perceived level of service being provided by the road. Through different types of equations, these levels of service are related to flows of different vehicles, e.g. St John and Glauz (1976) and Demarchi and Setti (2003). This method does

not relate perceived level of service to actual journey times and, thus, does not use an objective measure of PCE.

The headway method uses a measure of the relative road space occupied by a vehicle as a measure of the effect on passenger cars, e.g. see Werner and Morall (1976). Canagin and Chang (1984) suggested that more trucks increases the headway of trucks and, thus, the estimated PCE. Krames and Crowley (1986) suggested that this method is most or perhaps only appropriate for steady state conditions. Consequently, it may not be appropriate to use in changing conditions as is the case on many urban roads and expressways.

The queue discharge flow method is based upon the idea that exit from a queue is slower when the exiting vehicle is larger and this lead to passenger equivalences of greater than one, e.g. see Al-Kaisy et al (2002). This procedure is typically used with traffic simulation and an optimisation procedure, so in this sense is not estimated directly from field data.

Van Aerde and Yagar (1983) made a linear regression of a measure of speed on the flows of different vehicles. Chandra and Sikdar (2000) and Rahman and Nakamura (2005) have also used somewhat similar analyses that focus on relative speeds or different vehicles. This method directly relates speed to flows of different vehicles and is favoured in this thesis. However, it is important to allow for interactions between different vehicle flows and non-linear effects which have both been seen to be important, e.g. see Cunagin and Chang (1982) and Demarchi and Setti (2003). It is shown in Chapter 5, that interaction effects and a cumulative effect in the cost of congestion make this method more complex in theory and in empirical investigation.

A related method was used by Fan (1990) to consider PCEs in congested circumstances on Singapore expressways. The road volume to capacity measurement was linearly regressed on different vehicle flows and the regression coefficients interpreted as PCEs. This method ignores interaction and non-linear effects. The reported PCEs for larger vehicle types were higher than might be expected and close to but all less than three.

The Walker method used in the Highway Capacity Manual (1965) and later editions considers larger vehicles causing delays to passenger vehicles because of difficulties in overtaking. This method usually includes field observations of overtaking. Thus, the method is evidence based but does not consider directly speed reductions caused by larger vehicles.

The method referred to as vehicle hours considers the impact of trucks across intersections and arterial roads. The analysis is an extension of the level of service approach to include road sections and intersections, e.g. see Sumner at al (1984). In that the level of service is not objective and there is not an explicit measurement of journey time, this method can be criticised.

The relative travel time method measures PCEs by the ratios of travel times of different vehicles to the base (passenger) vehicle over urban networks, e.g. see Keller and Saklas (1984). This method has the merit of directly observing travel time but these ratios may reflect the slower speeds of larger vehicles rather than the impact of larger vehicles on passenger car speeds.

There are very many studies both academic and official that use PCEs to convert the numbers of different vehicle types to number of passenger cars for example, Smeed, (1964), Newbery (1987 and 1989; Mayeres et al. (1996); De Borger et al.

(1996); Sansom et al. (2001); Santos (2005); Highway Capacity Manual (2010), WEBTAG provided by Department for Transport (2011a) and COBA11 provided by Department for Transport (2012a).

It is important to consider the factors that may determine PCEs. The relative congestion impact of large vehicles varies with road conditions. For example, the discharge flow from a bottleneck is likely to involve a slower adjustment by large vehicles than passenger cars and a relatively greater impact on congestion, see Al-Kaisy (2006) and Elefteriadou, Torbic and Webster (1997).

A number of studies discuss the processes that result in the impact of trucks on passenger car speed, for example see Krammes and Crowley (1986); Al-Kaisy et al. (2002); Lindsey (2009) and Verhoef and Rouwendal (2003). According to Krammes and Crowley (1986), the presence of heavy vehicles causes physical and psychological effects on the other vehicles and their drivers respectively. In terms of physical effects, Al-Kaisy et al. (2002) suggests that the larger dimensions and the poorer operational characteristics of heavy vehicles compared to passenger cars are important in determining the effects on other traffic. Likewise, Lindsey (2009) suggests that trucks do not only use more road space but their larger size causes a greater degree of “blind vision” to other cars. In addition, he argues that trucks have poorer performance than a car either in accelerating or decelerating and these results in a greater impact on traffic out of steady state equilibrium. The larger size and poorer performance of trucks are likely to induce psychological responses from passenger car drivers. Correspondingly, Verhoef and Rouwendal (2003) argue that the speed adjustment made by each driver is endogenously determined by the presence and behaviour of other drivers on the road to adapt to risk of accidents. In addition, Lindsey

(2009b, p. 5) makes a similar statement saying that “the behavior of car drivers is affected by the presence of trucks in way that can affect safety”. He suggests that because of the psychological discomfort of car drivers from the presence of trucks, car drivers are likely to overtake trucks more quickly than cars. In addition to the safety concern, Al-Kaisy et al. (2002) suggest that the presence of trucks on the freeway results in car drivers keeping longer gaps both in front and behind trucks. As a result of this behavior, the speed of vehicles on the adjacent lane is affected. In contrast with Al-Kaisy et al. (2002), Vickrey (1963) suggests that car drivers are concerned about the close presence of trucks and will be reluctant to overtake when the space between their vehicle and a truck is shorter than a safe level. As a result of these circumstances and in congested conditions, car speeds are forced to be at truck speeds.

As noted before, many (but not all) of the studies using the methods previously discussed, engineering researchers have developed simulation models to estimate PCEs and increasingly sophisticated methods have incorporated variables such as traffic conditions, gradients, characteristics, vehicles’ physical and operational performance (see Werner and Morrall, 1976; St John, 1976; Huber, 1982; Keller and Saklas, 1984; Sumner, Hill and Shapiro, 1984; Roess and Messer, 1984; Krammes and Crowley, 1986; Fan, 1990; Lam, 1994; Elefteriadou, Torbic and Webster, 1997; Benekohal and Zhao, 2000; Demarchi and Setti, 2003; Al-Kaisy, Jung and Rukha, 2005; Al-Kaisy, 2006; Chitturi and Benekohal, 2007 and Geistefeldt, 2009). The issue here is that many engineering studies use simulation exercises that use different ideas about the congestive effect of different types of vehicles. The different ideas about measuring congestive effect lead to different approaches. It is suggested that the marginal external congestion cost of different

types of vehicles should be the basis of new studies as economic planning requires such a measure. It should be noted here that the there is evidence of non-linear and interactive effects of the congestion effect of different vehicle flows, e.g. see Arasan and Krishnamurthy (2008) and Al-Kaisy, Jung and Rakha (2005). This is the goal of the theoretical and empirical investigation in Chapter 5.

## 2.10 Value of Time

The marginal external congestion cost consists of two components, i.e. the value of time and the marginal external time cost as shown in (2.4.5).

$$MECC = b\left(\frac{-F}{V^2}\ \frac{dV}{dF}\right) \qquad (2.10.1)$$

where $b$ - Value of time

$V$ - Traffic speed, kilometres per hour

$F$ - Traffic Flow, vehicles-kilometres per hour

However, let us make clear that our thesis examines only the marginal external time cost $\frac{-F}{V^2}\ \frac{dV}{dF}$ but not the value of time $b$. This is because the purpose of the thesis is to model the physical congestion externality rather than its economic value. However, the value of time is a crucial parameter determining the marginal external time cost for road congestion pricing policy appraisal. Therefore, it is important to review the concept here.

The basic idea of value of time is discussed in the theory of time allocation (Becker, 1965). The conventional model of value of time simply assumes that an

individual has freedom to allocate time between various purposes, e.g. leisure (non-productive activity) and working (a productive activity). In appraisal of transport projects, time saving is considered as the most important user benefit (Hensher, 2011). However, unlike the technical speed flow relationship discussed previously, the value of time saved usually comes from an analysis of travellers' economic behaviour through experiments or observation of travel arrangements.

A simple approximation for the value of time for leisure is its opportunity cost which could be taken as the wage rate (Gwilliam, 1997). However, this ignores that people cannot really control their number of working hours. Therefore, it is suggested that the estimation of the value of time should incorporate the utility or disutility of work. As a result, the value of time can be higher or lower than the wage rate (Johnson, 1966; Oort, 1969 and Evans, 1972). For instance, the activity based model of value of time is developed by taking into account differences between valuation of working and non-working travel times (DeSerpa, 1971; Hensher, 1977; Calfee and Winston, 1998 and Mackie, Jara-Diaz and Fowkes, 2001 and De Palma and Lindsey, 2004), the idea of substitution of travel for other activities by rescheduling of preferred departure schedules (Small, 1982 and Jara-Diaz, 2003), and variation in the value of time spent on alternative traveling modes, e.g. time spent on trains and passenger cars.

Economists explain that an individual makes a decision to allocate time for specific activities and trades off this time with other resources. For instance, if drivers prefer to travel from work to home arriving at their destination on time, an increase in travel time will cause a disutility. In response to this problem, the driver has to reschedule his departure time and perhaps accept part of the increase in travel time. However, his decision depends on how he values the time saving from

rescheduling. Accordingly, the decision depends on the valuation of time for heterogeneous units of time, i.e. time for being in the car or being at work. (Small et al., 1999, Mackie, Jara-Diaz and Fowkes, 2001, Brownstone and Small, 2005; De Palma and Fosgerau, 2011 and Hensher, 2011)

Therefore the valuation of time models are generally made on various assumptions, for example the travel time (working or non-working time), relative values placed on various aspects of travel (i.e. walking time, waiting time, and travelling in-vehicle), length of distance, income variation, travel modes (car, bus, rail or freight) and variation of value of time over time. It is important to note that Wardman (1998) and Metz (2008) suggest that many of the studies of value of time are carried out through experiment rather than real world observation of actual behaviour in transport decisions.

Transport value of time studies are typically conducted using either disaggregated choice of stated preference or revealed preference data. For example, choices are analysed between modes or between alternative routes with the same mode with different time and money characteristics (Wardman, 2001). Estimation of willingness to pay for choices that reduce travel time are disaggregated behavioural parameter studies which can be conducted using stated or revealed preference data (Calfee and Winston, 1998; Lam and Small, 2001; Hensher, 2001a; Small et al.,2005). Metz (2008) notes that revealed preference and stated preference data are standard methods for modelling of value of time. The data used in revealed preference are alternative travel choices involving different costs whereas the stated preference uses hypothetical choices made by individuals of routes and modes involving different costs.

Wardman (1998) investigates a number of value of time studies in the UK. His review shows that the estimated values of time vary different greatly with circumstances and method of estimation. For example, diiffernces are revealed between studies that consider the area of the journey (urban, suburban, or inter-urban); the type of data; journey purpose; the choice context; the mode used; the , the purpose of the study; the means used to measure the stated preference; and whether logic or other checks were applied to the data. In the UK, the COBA guidelines provided by Department for Transport states that the estimation of the value of time should be weighted averages taking into account vehicle type, vehicle occupation, trip purpose, average wage, value of leisure time, etc ( Department of Transport (2012b), De Palma and Lindsey, 2004, and Newbery and Santos, 2002).

Many empirical studies of value of time have been carried out at the national level, for example Great Britain (Hague Consulting Group and Accent, 1996), the US (Walters II, 1995), the Netherlands (Hague Consulting Group, 1990 and Gunn and Rohr, 1996), Norway (Ramjerdi et al.,1997) Sweden (Alger et al.,1996) and Finland (Pursula and Kurri, 1996). However, that estimation of local values of time is important as there are a number of conditions that may vary locally and affect the aggregate value (McIntosh and Quarmby,1970 and Wardman, 1998).

In conclusion, we report in Table 2.10.1 a list of studies that estimate the value of time. The studies were carried out with different travel modes (i.e. car, bus, truck, rail and multimodes), different trip purposes (i.e. commute, leisure and commercial) and in various countries across the world (i.e. the USA, the UK, New Zealand, Germany, Switzerland, Chili, Spain and Australia). The research methodologies used for the valuation are stated and revealed preference, and use

weighted wage rate and mileage weighted income. Obviously, the results show a great variation of value of time for car mode between £4.02 to £28.92 per vehicle-hour and the values of time for public transportation (i.e. bus and rail) vary between £1.54-£18.09 per vehicle-hour. The value of time for commercial truck and multimodes are £40.45 and £8.47-£15.53 per vehicle-hour , respectively. The value of leisure time are £2.05-£18.54 per vehicle-hour. These large differences indicate that different possible values of time have an important quantitative impact on the possible marginal external congestion cost.

In conclusion, the value of time is important as the final component in estimating the marginal external congestion cost. This impact on the estimation of the degree of the externality means that it is important to carefully estimate the values of time. It is important to emphasize the great variation of value of time over various factors, e.g. heterogeneity of trip time, road users' behaviour, alternative transportation modes and research methodologies. This variation and the different methods of valuation show how difficult and important it is to value time correctly in estimating congestion externalities.

## 2.11 Conclusion

This literature review is concerned with the development of theoretically sound models of road congestion that can be used to estimate the costs of congestion. Unfortunately, this subject area is surrounded by unresolved debates and important problems.

## Table 2.10.1 Empirical Estimations of the Value of Time

| No. | Author | Mode | Data Collection Period | Location | Trip Purpose | Method | Value of Time ( per vehicle-hour) | Convert to 2012 price (£) |
|---|---|---|---|---|---|---|---|---|
| 1. | Small et al. (1999) | Car | 1995 | USA | Commute | SP | US$3.9 | 4.02 |
| 2. | Hensher (2001a) | Car | 1999 | NZ | Commute | SP | NZ$8.7 | 4.17 |
| 3. | Lam and Small (2001) | Car | 1997-1998 | USA | Commute | RP/SP | US$22.9 | 20.71 |
| 4. | Brownstone et al. (2003) | Car | 1997-1999 | USA | Commute | RP | US$30 | 26.88 |
| 5. | Brownstone and Small (2005) | Car | 1999-2000 | USA | Commute | RP/SP | US$12.6 | 11.46 |
| 6. | Small et al (2005) | Car | 1999-2000 | USA | Commute | RP/SP | US$21.5 | 19.56 |
| 7. | Steimetz and Brownstone (2005) | Car | 1999 | USA | Commute | RP | US$30 | 27.21 |
| 8. | Asensio and Matas (2008) | Car | N/A | ESP | Commute | SP | €14.1 | 12.2 |
| 9. | Patil, S. et al. (2011) | Car | 2008 | USA | Commute in urgent situation | SP | US$7.4-47.5 | 4.5-28.92 |
| | | | | | Commute in ordinary situation | SP | US$7.4-8.6 | 4.5-5.23 |
| 10 | Department for Transport (2011b): WEBTAG | Multi-mode | 2002 | UK | Commute | Weighted-wage rate | £11.28/hour | 15.53 |
| 11. | Department for Transport (2012): COBA11 | Multi-mode | 2000 | UK | Commute | Mileage-weighted income | £9.30/hour | 13.25 |
| 12. | Li, Hensher and Rose (2010) | Car | N/A | AUS | Commute | SP | AU$23.83-32.97 | 12.62-17.46 |
| | | Bus | | | | SP | AU$2.92-4.04 | 1.54-2.14 |
| 13. | Hollander (2006) | Bus | 2004 | UK | Commute | SP | £4.2/hour | 5.46 |
| 14. | Batley and Ibanez(2012) | Rail | 2007 | UK | Commute | SP | £15.4/hour | 18.09 |
| 15. | Bhat and Sardesai (2006) | Multi-mode | N/A | USA | Commute | RP/SP | US$12.2/hour | 8.47 |
| 16. | Smallkoski and Levinson (2005) | Truck | 2003 | USA | Commercial | SP | US$49.42/hour | 40.45 |
| 17 | Jara-Diaz et al. (2008) | NA | 2002 | DEU | Leisure | SP | US$26.7 | 24.52 |
| | | | 2004 | CHL | Leisure | SP | US$2.9 | 2.05 |
| | | | 2005 | CHE | Leisure | SP | US$26.7 | 18.54 |

Note: (1) SP stands for stated preference. (2) RP stands for revealed preference. (3) Multimodes include car (drive alone and shared ride), car, bus and rail in their SP. (4) 2012 price (£) is the value are all converted into £ with the annual exchange rates and inflated to 2012 based on CPI.

The standard conventional speed flow analysis has the problem of being a static model and the derivation of the marginal external congestion cost is based upon additional flow rather than the more realistic additional vehicle or density. In this model, the congestion cost comes from journeys taking longer. In a one period model, there are no additional periods in which to start the journey earlier or complete the journey later. In addition, the standard static model takes no account of the external costs coming from other road users being denied their preferred scheduling journey start and completion times. It is suggested that the solution to some of these problems can be provided by a density based multiperiod model.

The standard speed flow analysis fails to provide a theory of hypercongestion that allows estimation of marginal external congestion costs. In addition, the subjects of hypercongestion equilibria, stability and optimality have resulted in much argument which does not appear to be resolved. It is suggested again that the solution to some of these problems can be provided by a density based multiperiod model.

Finally, previous mainly engineering studies of the congestion effects and impacts of different types of vehicles have not been based upon a sound economic model. The calculation of these effects have used simulations rather than estimation based upon field data. It is suggested that a more economic theory based approach using field data is required to estimate the impacts of different types of vehicles.