Speed, flow and density of motorway traffic

S O Notley, N Bourne and N B Taylor
Speed, flow and density of motorway traffic

S O Notley, N Bourne and N B Taylor

TRL Insight Report INS003
CONTENTS

Executive summary v
Abstract vi

1 Introduction 1

2 Background 2
   2.1 Highways Agency TRaffic Information System 2
   2.2 Observed speed–flow relationships 2
   2.3 Traffic flow theory 3

3 Analysis of relationships in traffic data 5
   3.1 Features of speed–density relationships 5
   3.2 Traffic phases in the speed–flow relationship 5
   3.3 A simple interpretive model 6
   3.4 Phase transitions and flow breakdown 6
   3.5 Traffic phases in the flow–density relationship 7
   3.6 The problem of measuring capacity 7
   3.7 Extended motorway queues and shockwaves 9
   3.8 Traffic in the uncongested phase 12

4 External influences on the relationships 13
   4.1 Scatter in the speed–flow relationship 13
   4.2 Variation in relationships by day of week 14
   4.3 Time of day and heavy goods vehicle proportions 14
   4.4 Summary 17

5 Modelling the speed–flow–density relationship 17
   5.1 Connections between speed and density 17
   5.2 Stimulus-response car-following models 18
Executive summary

In recent years, the UK has seen a shift in emphasis away from road building and towards the use of technology to fully utilise the capacity of existing roads. In order to develop technologies that effectively optimise traffic flow, it is vital to understand the underlying processes that determine the speed and flow of traffic.

It seems intuitively obvious that the more traffic that tries to use a given section of road, the slower it must move, but the precise mechanisms behind this relationship are surprisingly elusive. Since the 1930s, when the first freeways appeared in the USA, and most intensively from the late 1960s to the 1980s, when congestion became a regular occurrence, traffic measurements have been studied and theories developed. The early theorists sought basic principles and simple models, in the best tradition of physics. As the “classical” models proved inadequate in various ways, the emphasis switched to empirical models, such as those developed for the Department for Transport’s Cost-Benefit Analysis computer program in the UK and for the Highway capacity manual in the USA. More recently, studies have drifted towards the academic, but these may risk trying to read too much into too little data, or being too complex and detailed to be useful to practitioners.

The availability of large amounts of data collected automatically raises the possibility of validating theories and models. Examples of such data support the concept of a two-value speed–flow curve where a given flow can be sustained at two different speeds, converging at a single optimum speed and flow (Section 2). High-resolution motorway data support the theory that these two speeds correspond to congested and free-flowing states (Section 3). The individual measurements of speed and flow stored in the Highways Agency TRaffic Information System show considerable scatter — with a large range of speeds observed for any given density. It is shown that some of this scatter can be attributed to the dependence of the relationships themselves on external influences such as traffic composition (Section 4).

If traffic flow is to be modelled accurately, it may be necessary to determine how such external influences affect the parameters of the relationships. The data support the idea that average speed decreases with increasing traffic density, but only above a certain density. Below this, speed is virtually independent of density. This is not a feature of the “classical” macroscopic models, suggesting that none of them fully describes traffic over the full range of possible densities, and other properties raise issues, in particular the ability to support shockwaves (Section 5).

The processes governing the transition between these states in unobstructed traffic are not fully understood. However, the transitions in the presence of a bottleneck are consistent with deterministic queuing theory, and suitably enhanced methods can be used to model major motorway queues, including those produced by moving bottlenecks (Section 6).

While the increase in monitoring technology has led to improvements in modelling and the development of visualisation tools such as Motorway Traffic Viewer, there is still considerable scope for further research into the fundamental processes that drive traffic flow. The huge amounts of data now available present a previously unavailable opportunity to thoroughly test and refine theoretical models. Models based on individual driver behaviour have proved difficult to calibrate because they depend on internal variables that are not directly measurable in aggregate traffic. Macroscopic models, by and large, can be calibrated from aggregate measurements, so in this regard may be more useful, but still need to be shown to have merit as physical explanations, not just as convenient formulations. The work described in this Insight Report is beginning to tease out the properties of different models, and to identify what needs to be measured and what essential features model should possess. These are key points that must be addressed if a model is to be useful and meaningful across all applications. The matter will not be settled by either theoretical model formulation or data fitting alone, but will need the convergence of both.
Abstract

It seems intuitively obvious that the more traffic that tries to use a given section of road, the slower it must move, but the precise mechanisms behind this relationship are surprisingly elusive. The availability of large amounts of data collected automatically raises the possibility of validating theories and models. This Insight Report examines the features of some actual data and speed–flow–density relationships, and “classical” models of speed, flow and density in the context of the wealth of detailed traffic data now available. Data from detectors support the idea that average speed decreases with increasing traffic density, although the data suggest that this decrease is only significant above a certain density. Below this, speed is virtually independent of density. This is not a feature of “classical” macroscopic models, suggesting that none of them fully describes traffic over the full range of possible densities. The role of speed, flow and density in queuing theory is also examined, including a case study of modelling a moving bottleneck.
1 Introduction

It seems intuitively obvious that the more traffic that tries to use a given section of road, the slower it must move, but the precise mechanisms behind this relationship are surprisingly elusive. Small and Chu (1997) make the point that an additional vehicle cannot confer a positive externality on existing traffic, but in congestion this seems to be violated since speed appears to increase with flow, resulting in an overall two-value relationship. Since the 1930s, when the first freeways appeared in the USA, and most intensively from the late 1960s to the 1980s, when congestion became a more regular occurrence, traffic measurements have been studied and theories developed. The early theorists sought basic principles and simple models, in the best tradition of physics. As these “classical” models proved inadequate in various ways, the emphasis switched to empirical models, such as those developed for the Department for Transport’s (DfT) Cost-Benefit Analysis (COBA) computer program (Highways agency et al., 2002) in the UK and for the Highway capacity manual (Transportation Research Board, 2000) in the USA. These enabled operational and economic appraisals to be conducted.

In recent years, the “fundamental” relationship between flow, speed and density has been questioned. Kerner (2004) has identified synchronised flow on multi-lane roads in which flow and density appear to be unrelated. In seeking to explain flow breakdown and “shockwaves”, researchers have variously used detailed driver behaviour, continuum theory, phase changes or catastrophe theory, and interactions between different driving populations. However, academic approaches may risk reading too much into too little data, or relying too much on internal behavioural variables, which are difficult to measure.

What is new about the present time is the massive increase in the amount of traffic data available, thanks to the continuing roll-out of automated data-gathering technologies on UK roads. Much of the data gathered from such systems are consolidated in the Highways Agency Traffic Information System (HATRIS) database, where they are available for analysis. This means that it is now feasible to validate theoretical models against large numbers of observations without the need for costly or time-consuming traffic surveys.

The focus of this report is two-fold:
- To provide a broad overview of existing theories of traffic flow.
- To examine the results and implications of these theories in the context of speed and flow data available from HATRIS.

By providing a holistic view of the theory surrounding traffic flow and placing this in context using real-world observations, it is hoped that a clearer understanding can be gained of the fundamental nature of traffic flow and of the properties a theoretical model must have in order to reproduce this. Traffic management is becoming increasingly sophisticated, with systems capable of setting signs and signals in response to traffic conditions or at the request of an operator. If these technologies are to be exploited to their full potential, it is important that designers and operators are equipped with an ability to understand and predict traffic flow and how it will respond to their actions. This Insight Report aims to contribute to that understanding and lay the foundations for further research into accessible, functional traffic flow models.

This report concentrates solely on motorways for a number of reasons. As motorways have very few entry and exit points and no roundabouts or main-carriageway junctions, they closely resemble the idealised “roads” used to develop flow theories. The fact that motorways have separate carriageways and controlled alignment means that complex effects of road geometry and overtaking against opposing traffic need not be considered. In addition, the highest-quality data are typically available from detector systems such as Motorway Incident Detection and Automatic Signalling (MIDAS), currently only installed on motorways.

This Insight Report is not intended to be an exhaustive review of the entire field of flow theory. However, a list of further reading is provided for further information, including some extensive reviews.
2 Background

2.1 Highways Agency TRaffic Information System

HATRIS is a system of databases and procedures that allow the Highways Agency (HA) to process and store the wealth of available traffic data from the primary network for which it is responsible. Central to HATRIS is the Merged All Data for Journeys (MADJ) table, which contains journey time and flow data for every 15 minutes on every junction-to-junction link of HA roads since September 2002. The data used for analysis in this study cover the three-year period from 1 January 2003 to 31 December 2005. The link codes (e.g., LM299, which corresponds to junctions 10–11 of the M25) are the HATRIS link references. The MADJ table combines journey time (and thus average speed) data from up to five different sources:

- MIDAS data are gathered by inductive loops embedded on the road. MIDAS consistently collects the traffic data and has a large sample size, but is currently rolled out on only a fraction of the motorway network and does not cover any non-motorway links.
- Trafficmaster Automatic Number Plate Reading (ANPR) data are predominantly used on non-motorway links and cover a large proportion of the network.
- ITIS (from ITIS Holdings PLC) data are gathered from GPS systems installed principally in commercial vehicles, giving good network coverage, but meaning samples are unlikely to be representative. Sample sizes are also typically quite small (about five vehicles per hour). As such, ITIS data are only used in MADJ where other data are not available, principally on non-MIDAS motorway links. These data are no longer provided to HATRIS so no data are available after December 2006.
- Trafficmaster GPS data are also gathered from GPS systems. This source replaced ITIS in January 2008, although some data are available from July 2007 onwards. This data source was not available when the analysis described in this report was originally conducted.
- National Traffic Control Centre (NTCC) data are gathered by a relatively sparse network of ANPR cameras.

Journey times from each of these sources are included separately (where available) and also as a “merged” journey time, where a number of criteria are used to select the most reliable source for each 15-minute period. In this report, only MIDAS data are used, as MIDAS is believed to be the most reliable data source for motorways. HATRIS includes algorithms to estimate the speed and flow on a link when observed data are not available. The HATRIS data used in this report have been filtered to include only observed speed and flow data.

2.2 Observed speed–flow relationships

Figure 2.1 shows a typical plot of hourly average speed and flow on a motorway link, covering a three-year period. On links that experience recurrent congestion, as in Figure 2.1, observed speed–flow relationships often cluster around two values of speed for each value of flow, and observed maximum flows are generally similar on different links with the same number of lanes. It is tempting to infer that the maximum flow represents the inherent “capacity” of the road section. This can be the case only if the flow is free and not limited by external influences acting outside the section, e.g., a nearby bottleneck.

While the relationship shown in Figure 2.1 is broadly typical for a motorway link, some variation is observed between links.

*Flow data in the MADJ table are sourced from the TRAFFIC flow Data System (TRADS) and are provided with a resolution of one hour, meaning that all four 15-minute periods within the hour will be assigned the same flow.
In order to estimate journey times and throughputs, it is necessary to understand how this characteristic shape arises, what are the causes of variability and what factors dictate where in the observed distribution the traffic state at a given time will fall.

2.3 Traffic flow theory
Traffic flow theory attempts to relate measurements of traffic flow, speed and concentration. An extensive discussion of the definition of each of these quantities is provided by Professor F. Hall in Chapter 2 of Gartner et al. (1992):

- Flow \( q \) is defined as the number of vehicles passing a point during a given time period.
- Speed \( u \) can be measured either as a time-mean, measured at a single point, or a space-mean, representing an average of instantaneous vehicle speeds on a section.
- Concentration can be defined in terms of occupancy or density. Density \( k \) is the number of vehicles present per unit length of road at a given moment, while occupancy measures the proportion of time that a vehicle is present over the detector.

All of these variables can vary with both time and in position along the link (Wardrop, 1952; Gerlough and Huber, 1975). The fundamental stream equation (Equation 2.1) relates the three variables of flow, speed and density for any conserved body of traffic whose extent in time and distance can be defined unambiguously, including small time and distance intervals:

\[
q = uk
\]

(Equation 2.1)

Treating the variables as fields defined at points in space and time, Equation 2.1 represents a surface in a three-dimensional phase space for potential traffic states (Gartner et al., 1992). Once an operational relationship is specified between any pair of the variables, Equation 2.1 determines the relationships between the other two pairs.

One can visualise these by plotting projections of the overall relationship on the \( u-k \), \( u-q \) and \( q-k \) planes. However, it is not generally possible to hold one variable fixed while varying another in order to observe the effect on the third (Papacostas, 1987), except in certain dynamic cases, e.g. where a constant flow of traffic is accelerating away from a queue.

The simplest pairwise relationship is the linear relationship proposed by Greenshields (1935) between speed and density, \( u-k \) (Equation 2.2), where \( u \), is maximum or free speed and \( k_j \) is maximum or jam density:

\[
u = u_f \left(1 - \frac{k}{k_j}\right)
\]

(Equation 2.2)

This can be taken to imply that drivers choose their speed according to the density of traffic around them, with a theoretical maximum (possibly “desired”) speed \( u_f \) at zero density, dropping linearly to zero speed when density reaches jam density \( k_j \). Figure 2.2 shows that if this is assumed, parabolic relationships follow between density and flow, and speed against flow, the speed–flow relationship being two-valued. In each panel of Figure 2.2, the dashed background lines represent contours of the third variable, so in the first panel one can see the lines of constant flow, which are simply given by the functional family (using Equation 2.1).

Another important relationship in traffic is given by Equation 2.3, which expresses the conservation of vehicles at a boundary where the traffic state changes, subscripts \( dn \) and \( up \) representing the downstream and upstream states, respectively, and \( u_w \) being the speed at which the boundary propagates (positive meaning downstream):

\[
u_w = \frac{q_{dn} - q_{up}}{k_{dn} - k_{up}}
\]

(Equation 2.3)

This relationship is particularly relevant to queue formation as discussed in Section 6.
Figure 3.1 Speed–density plot for the M25 motorway link LM299

Figure 3.2 Logarithmic speed–density plot for the M25 motorway link LM299 (compare Figure 3.1)
3 Analysis of relationships in traffic data

3.1 Features of speed–density relationships
The speed–density plot corresponding to Figure 2.1 is shown in Figure 3.1. These density data are calculated from HATRIS speed and flow data using Equation 2.1.

Figure 3.1 shows a drop in speeds at the lowest densities, corresponding to a similar feature at the lowest flows in Figure 2.1. It is believed that this is caused by a high proportion of heavy goods vehicles (HGVs) overnight rather than being a function of density. This is discussed further in Section 4.3. Where density is less than about 50 veh/km (on three lanes), speed appears to be broadly constant. Above this density, speed decreases with increasing density, albeit not in the linear fashion proposed in Section 2.3.

The density scale in Figure 3.1 concentrates many points at higher speeds and lower densities. Figure 3.2 shows the same data as Figure 3.1 on a logarithmic density scale, supporting the idea that the speed–density curve can be split into two separate and approximately linear (on these axes) halves above and below a “critical” density, corresponding to the upper and lower halves of the speed–flow relationship shown in Figure 2.1. This finding is consistent with much of the literature, including Gartner et al. (1992), Kockelman (2001) and Kerner (2004). An intuitive interpretation is that at densities below the critical value, vehicles have sufficient separation and freedom of manoeuvre for their speeds to be little affected by the presence of other vehicles, whereas above it obstruction by nearby vehicles begins to limit their speed. The low-density regime is also affected by speed limits.

3.2 Traffic phases in the speed–flow relationship
In speed–flow plots, data are commonly concentrated in two regions, separated by a gap in which points are more sparsely distributed; these are highlighted in Figure 3.3.

These regions can be interpreted as representing two distinct “phases” of traffic, shown schematically in Figure 3.4. The upper region can be described as “freely flowing” or “uncongested” traffic, while the lower region represents “congestion”. The distinction between the two phases is instinctively understood by motorists and transport engineers alike, but is not trivial to describe mathematically. Daganzo et al. (1999) define free flow more rigorously as a state in which disturbances can only travel downstream and queuing (congestion) as a state in which disturbances travel upstream. However, the direction in which disturbances move depends on the slope of the flow–density relationship (which is consistent with that of the speed–flow relationship). So, if

---

3 ANALYSIS OF RELATIONSHIPS IN TRAFFIC DATA

---

Figure 3.3 Traffic phases in LM299 data

Figure 3.4 Two phases of traffic states
one formula governs all or part of both regions, as is possible for several of the macroscopic formulae discussed later in Section 5, the distinction between free-flowing and congested phases, as opposed to traffic states, becomes less clear.

3.3 A simple interpretive model

A natural step from Figure 3.2 is to propose a simple model in which the uncongested and congested parts are approximated by straight lines passing through three key points – “free-flow” speed, the “critical” point and a point representing “jam” conditions – as illustrated by Figure 3.5 (following suggestion by G Skrobanski, Highways Agency).

From inspection, we can make the simplifying assumption that \( u_f = u_c \), so the model is:

\[
\begin{align*}
  u &= u_c - \frac{(u_c - u) \ln(k/k_c)}{\ln(k_j/k_c)} & (k > k_c) \\
  u &= u_c & (k \leq k_c)
\end{align*}
\]

(Equation 3.1)

In addition, the jam speed \( u_j \) acts as a small displacement to all the speeds without changing the form of the function. If it is taken to be zero, then the formula for congested traffic simplifies to that of Greenberg (1959):

\[
  u = u_c \ln \left( \frac{k_j}{k_c} \right)
\]

(Equation 3.2)

Certain features of this and related models are discussed in Section 5.

3.4 Phase transitions and flow breakdown

Dynamical plots of speed flow, such as Figure 3.6, which plots MIDAS’ data colour-coded by time of day, show that at times of lower demand the traffic state remains in the upper region, but as demand increases there comes a point when speed and flow both drop and the traffic enters a congested state in the lower region. The traffic remains in the lower region until demand drops again (ie until the end of the peak), at which point the speeds recover. However, this occurs at lower flow, so the traffic state “cuts across” from the lower phase to the upper. The process by which free-flowing traffic first passes into the congested regime is often referred to as “flow breakdown”. It is not clear whether Figure 3.6 shows the initial occurrence of flow breakdown and recovery or simply a queue forming and then dissipating due to an incidence of flow breakdown further downstream. The transition cycle is illustrated schematically in Figure 3.7.

In Figure 3.6, the speed–flow state varies constantly, even in the absence of any apparent external influence, and the link does not appear to operate at a well-defined capacity for any significant period of time. Indeed, the highest flow measured, around 2800 veh/h/lane, is well above what is usually taken to be the maximum flow on a motorway, ie 2100–2300 veh/h/lane (Hunt et al., 1991; Hunt and Yousif, 1994; Taylor, 2006a). However, it is important to bear in mind that these data come from MIDAS loops at set locations on the link, and as such only represent spot-speeds and flows. The maximum flow recorded over a short period may not be sustainable; in fact, “capacity” is really only useful as an average measure (see Section 3.6).

White and Abou-Rahme (TRL Unpublished Project Report UPR/T/140/02, 2003) find that flow breakdown is a probabilistic phenomenon and that the probability of flow breakdown increases with the local transient flow. Flow

---

\*\* The MIDAS data used here are not from HATRIS. They are data from a single-loop detector at one-minute intervals prior to aggregation for HATRIS.
breakdown may be initiated by a single driver momentarily slowing to avoid a conflict. Local increase in demand may result from traffic merging, and the probability of occurrence may also depend on local geometric factors. In heavy traffic, such a disturbance propagating upstream can be amplified and queuing may result. Flow breakdown often occurs at static “seed points”, but the observable changes at those points can be slight. It is not possible to say from observing one site alone whether a queue has been caused by flow breakdown or by a physical bottleneck, whether drivers in the congested regime will continue to move at low speed for the entire length of the link or whether they will move out of the congestion further downstream.

3.5 Traffic phases in the flow–density relationship

There are a number of papers addressing the behaviour of regimes/phases and phase transitions in traffic, including Daganzo et al. (1999), Kockelman (2001), Helbing et al. (2002), Kamarianakis and Prastacos (2004), Kerner (2004) and Öğüt (2004). In these studies, the most popular representation of these two phases is in the \( q-k \) plane, as shown in Figure 3.8 for the motorway link LM299. The slope of each point’s position vector, with respect to the origin, corresponds to its speed, so the dense near-linear region contains points representing high speed, around 100 km/h, then starts to fall off to around a density of 50 veh/km, then rapidly loses coherence after flow breakdown sets in. This kind of picture gives greater credence to the idea that the processes governing the phases are fundamentally different.

3.6 The problem of measuring capacity

The difficulty of defining motorway capacity is illustrated by Figure 3.9, which shows the maximum and average maximum lane flows measured by all main carriageway MIDAS sites on the M60 in April 2006 as a function of the time “window” over which flow is measured.

The shortest possible measurement window is one minute, the resolution to which MIDAS data are reported. The longest window is a whole day (1440 minutes). For longer windows (> 100 minutes), it is likely that demand is insufficient for the motorway to remain saturated over the whole window, meaning that the measured maximum flow is more likely to be a measurement of demand than capacity. An interesting
feature of Figure 3.9 is that maximum average flow declines rapidly with increasing window size even for the shortest windows. This suggests that capacity measurements should include the measurement period. For example, the maximum flow observed in a one-minute window is equivalent to about 2800 veh/h, but the maximum flow sustained for a ten-minute period is only equivalent to about 2500 veh/h. Possible reasons for this effect are:

- For the shortest time windows, there may be averaging of stochastic variations.
- Over slightly longer periods, transient high flows may lead probabilistically to flow breakdowns, so that average flow is reduced.
- For long time windows, off-peak flows reduce the overall average, which no longer represents capacity.

Figure 3.8 Flow–density plot for the M25 motorway link LM299

Figure 3.9 Maximum lane flow against measurement “window” (logarithmic scale) for whole M60 motorway
3.7 Extended motorway queues and shockwaves

The Motorway Traffic Viewer (MTV) application developed at TRL enables speed data to be plotted using a grey scale on time and distance axes, showing clearly the variations in space and time, and making shockwaves visible. Figure 3.10 shows a plot of speed on junctions 10–12 of the M25 during the morning of 21 July 2008.

In the MTV plot, time increases to the right and distance along the link upwards. Lighter colours represent lower speeds, and vehicle trajectories proceed diagonally upwards and to the right. Of particular interest in Figure 3.10 are the shockwaves (or stop-start waves) visible as diagonal white streaks propagating upstream. This complex pattern of queuing occurs even though variable speed limits and queue protection may be in operation.

Each vehicle passing through the “queue” experiences a series of fluctuations in speed and density, even being brought temporarily to a halt. The waves originate from identifiable locations known as seed points, which are often associated with physical locations such as junction on-slopes. The seed points differ conceptually from bottlenecks in that they do not represent a capacity restriction greatly lower than demand on a medium-term (eg hourly) basis, but fall into one of two following categories:

- They have a capacity so close to the average incoming flow that a very slight fluctuation in the incoming traffic flow effectively causes capacity to be exceeded.
- They have some physical property that may cause capacity to momentarily reduce. For example, a motorway entering an uphill gradient can suddenly reduce in effective capacity when a HGV loses speed, or simply because vehicles unintentionally fail to maintain speed.

Figure 3.11 shows a seed point located near the Junction 15 merge (the merge location is labelled “J15(m)” and the seed point highlighted in red) on the M25 causing shockwaves to propagate upstream. This differs from the solid “block” of low speeds caused by queuing following an accident shown in Figure 3.12 (the accident is highlighted in red), although shockwaves begin to form within the queue once the obstruction is cleared and the queue is allowed to disperse.

When a seed point is triggered by a fluctuation in demand or reduction in capacity, a small region of slower-moving, denser traffic is formed and flow drops to match capacity, just as with a bottleneck. Figure 3.13 shows that this is almost imperceptible at what is considered to be a seed point – in this case, the Junction 14 merge (loop location address 4927A). Loops further downstream, at 4932A, 4935A and 4940A, show freely flowing traffic (the speeds are lower between about 15:00 and 19:00 because of variable speed limits). The loop upstream, 4922A, clearly shows evidence of congested traffic, while at seed point 4927A there is evidence of traffic in the process of recovery as it passes through the bottleneck.
SPEED, FLOW AND DENSITY OF MOTORWAY TRAFFIC

Figure 3.11 An MTV plot showing a seed point

Figure 3.12 An MTV plot showing queuing following an accident
For flow breakdown at a seed point to be sustained, the tail of the region must propagate upstream, as described in Section 6, so speed must have fallen sufficiently to drive a few vehicles temporarily into the congested traffic phase. Where the slope of the congested speed–flow relationship is positive and increasing, if constant flow is to be maintained then any fluctuation in speed requires average speed to increase and average density to decrease. Conversely, fluctuation around average speed or density will tend to cause a net reduction in flow.

The effect of fluctuations can be addressed microscopically – see, for example, Helbing (2001), which includes an extensive bibliography – but it is possible that it can be approached macroscopically. According to the Duncan/Banks speed–density model and the continuity equation 2.3, any disturbance in a congested stream propagates upstream at speed \( \lambda/T \), around 19 km/h in practice. In Greenberg’s model, the propagation speed is \( u_0 - u \), where \( u \) is the average speed of traffic. This fails at low speeds, but implies that propagation speed is greatest when the traffic speed is lowest. Papacostas’ model also has this property. If this is characteristic of actual traffic, then shockwaves, which cause traffic to stop even momentarily,
will catch up with and overwhelm any waves of lesser disturbance, explaining why they dominate the MTV plot.

It has been noted that shockwaves appear to be quite regularly spaced in time, with a roughly constant period elapsing between the formation of each shockwave at a given seed point (White and Abou-Rahme, TRL Unpublished Project Report UPR/T/140/02, 2003). Their proposed explanation is that traffic becomes less “smooth” the longer it remains in a free-flow state between regions of congestion. Traffic dispersing from the front of a shockwave must eventually pass through the seed point that originally triggered the wave. When the traffic leaves the wave, it is relatively smooth, having been forced into an orderly queue from which vehicles depart at roughly regular intervals. After some time in free flow, dispersion in vehicle speeds causes fluctuations to occur as faster vehicles catch slower ones and, although the overall flow remains the same, it is no longer homogeneous. As the shockwave propagates further from the seed point, the traffic arriving at the seed point will contain more severe fluctuations, until eventually one is strong enough to reactivate the seed point and the process starts again. The plot also shows that shockwaves tend to broaden with time, while the spaces between them appear darker in the plot, implying that traffic moves more quickly there.

Note that, although the large-scale shockwaves seen on MTV are typically associated with exogenous physical seed points, it is possible that, even on an idealised road with no seed points, shockwaves could form endogenously within traffic if at some point the fluctuation in flow exceeds the capacity of the road, or more concretely, the speed of at least one vehicle drops low enough to push the local traffic state into the congested regime. Endogenous shockwaves could also form behind events that momentarily reduce capacity, such as a car pulling onto the hard shoulder or a HGV performing an overtaking manoeuvre. However, Daganzo et al. (1999) argue that traffic flow can be fully understood without consideration of these endogenous waves. It has been suggested that the only way to be certain is to observe in detail the behaviour of a large number of individual vehicles in the vicinity of a seed point. This has not been achieved because the methods used (eg helicopter-mounted camera) have not been able to cover a long enough section of road simultaneously, and an endogenous seed point, by its nature, is bound to be elusive. In principle, microscopic simulation would be more economical, but no microscopic model has been proved sufficiently realistic.

3.8 Traffic in the uncongested phase
Traffic in the congested phase can be viewed as moving in a number of independent lanes, each operating at capacity, in which movement is governed by short-range interactions. Traffic in the uncongested phase has in principle a richer repertoire of potential behaviour, including considerable freedom to choose its speed, headway and acceleration, and to change lane or overtake. Interactions can take place over variable times and distances, making it less likely that local models such as those for the congested regime can be applied. Helbing (2001) provides graphical evidence that the difference in speed between vehicles in parallel lanes increases not only in magnitude but also in its variability at low densities. In addition to increasing mutual obstruction of vehicles as flow increases, speeds can be affected by “lateral discomfort” or “psychological friction” between adjacent lanes. Although little appears to be known about the effect of lane width, it may be supposed that narrow lanes increase such effects.

In COBA (Highways Agency et al., 2002), these complex effects are subsumed by a bi-linear model, where the overall speed–flow relationship has a zero or slightly negative slope up to a certain “break flow”, after which it changes to a steeper slope (COBA input actually allows for bespoke relationships with up to five linear sections). However, it is not known how much of the original data giving rise to the steeper slope at higher flows included points representing the onset of or recovery from congestion, or included traffic states affected by conditions upstream or downstream. COBA contains calibrations for a wide range of road types and geometric parameters, but is descriptive rather than explanatory, which somewhat limits its predictive ability.

Davidson (1966, 1978) attempted to model uncongested flow by a series of small queues, using steady-state queuing theory with modified parameters (see also the discussion by Akçelik, 1991). However, this has both conceptual problems (through stretching an analogy with an equilibrium queuing process, which becomes undefined when demand equals or exceeds capacity) and calibration problems (relying on a capacity parameter whose physical meaning is obscure). A model that depends on calibrating internal parameters is more difficult to justify than an empirical model that produces similar results, such as COBA.

Another type of approach is to suppose that vehicles have a desired speed or spacing, which is then constrained by interactions with other vehicles (eg Addison and Low, 1996). Again, this depends on internal parameters, which are not directly measurable.
4 External influences on the relationships

4.1 Scatter in the speed–flow relationship

The scatter plots in Section 3 earlier appear to show that a large amount of data appears in the lower portion of the speed–flow curve, implying congestion. However, the large number of points plotted means that the graph becomes saturated and it is impossible to see the true distribution. To better understand the distribution, one can plot histograms to show the frequency distributions of speeds, flows and densities. In order to analyse scatter on the two-dimensional plots, two-dimensional frequency distributions must be plotted. An example is given in Figure 4.1, which displays the frequency of points in speed–flow bins using a colour scale.

Data are analysed within the three-year period from 1 January 2003 to 31 December 2005. Speeds of zero are excluded and densities and flows are measured per lane. The bin dimensions are 1 km/h x 20 veh/h/lane. Colours are defined on a logarithmic scale because there are just a few bins with a very large count while most are within a small range of low counts. The logarithmic scale increases the definition at low counts.

It can be seen in Figure 4.1 that the vast majority of the data lies along the upper branch of the speed–flow curve, which generally corresponds to low-density, freely flowing or uncongested traffic. This finding is not unexpected, since the counts represent individual hours for the entire three-year period, and only a few hours in each day are congested. Figure 4.2 shows a different motorway link that is less congested.

The speed–density relationships plotted from HATRIS data show a degree of “scatter” with a relatively large range of speeds observed for a given density. It is feasible that this is simply a random variation due to different drivers being prepared to accept differing headways, or having different desired speeds under free-flow conditions. It is also possible that the observed scatter could be due to some wholesale change in the speed–density relationship dependent on one or more

![Figure 4.1 Two-dimensional speed–flow histogram for motorway link LM299](image1)

![Figure 4.2 Two-dimensional speed–flow histogram for motorway link LM784 (junctions 22–23 of the M60)](image2)
external factors. For example, time of day, time of year, weather or light conditions might all affect the collective behaviour of drivers and thus change the speed–density relationship. If the observed relationship is actually the sum of multiple subtly differing relationships, this could explain the observed scatter. Sections 4.2 and 4.3 address some possible causes of variation in the relationship between speed, flow and density.

4.2 Variation in relationships by day of week

It can be seen in Figure 4.3 that the day of the week appears to have little effect on the shape of the $u$–$k$ relationship for LM299, although there is large scatter particularly at low (free) flows on all days; perhaps more so on weekdays. The sloping high-density congested region appears to be the same for all days (although there is little congestion on weekends) but the low-density speed appears to be slightly higher on weekends, particularly Sundays. However, this should not necessarily be taken as evidence that drivers will each choose to drive faster simply because it is Sunday. The higher speeds are likely to be a result of a different group of drivers experiencing free-flow conditions on a Sunday than during the week.

The apparent dependence of free-flow speed on day type in the low-density data can be further analysed by plotting speed–frequency distributions for the lower densities by day-type groups (Figure 4.4).

In Figure 4.4, all data with density of less than 15 veh/km/lane are included and split into six “day group” categories. The chart therefore includes data from a range of flows, but as this portion of the speed–flow plot is flat, there should be little broadening due to the range of flows. The frequency distributions are normalised so that the count is a fraction of the total count for each day group category.

Figure 4.4 shows that the variation between day types partially accounts for the scatter in speeds in free flow, though the distributions in each day are still fairly broad. It is notable that the upper end of the distribution is dominated by weekend data and the lower end by weekdays (excluding bank holidays).

The peaks of the Sundays and bank holidays are roughly coincident, and at higher speeds than the peaks of all the weekdays. Saturdays appear to be somewhere in between, having one peak in line with that of Mondays and another in line with that of Sundays. It is also apparent that there are two peaks for some of the groups (particularly Tuesdays–Thursdays, Fridays and Saturdays), suggesting another division within each group into two peak speeds. This split could be due to another division of day types within each of these groups, perhaps between summer and winter or school term times and holidays, as accounted for in the HATRIS day types; or it could be due to some other cultural or seasonal effect. It could otherwise be a result of the wide range of densities and flows covered, although the expected effect of this is a broadening of the peak and not a split.

4.3 Time of day and heavy goods vehicle proportions

It was seen in Section 4.2 that the speed of traffic at low densities is notably higher at weekends. It was suggested that this was more likely to be the result of a change in the traffic demographic, rather than a direct consequence of the day of the week. HATRIS does not contain any information on the composition of traffic. However, MIDAS loops gather far more data than are included in HATRIS, including flow statistics for four different classes of vehicle. These data can be used to estimate HGV percentages on links with enabled MIDAS loops.
Using a subset of MIDAS traffic count data (in which flows are broken down into four length categories), the HGV percentage on LM299 was estimated for every hour of 2006. A section of this sample is shown graphically in Figure 4.5. It is clear from this figure that the HGV percentages follow a strong weekly pattern, peaking overnight and dropping during peak flow times and at weekends. It is also interesting to note that Sunday early mornings have a far lower percentage of HGVs than other overnight periods.

It is clear from Figure 4.5 that the HGV percentage tends to peak at times when flow is low. This is intuitively sensible: not only do HGV drivers not have the same motivations to travel at certain times as other drivers, but they will often actively choose not to drive when demand is high in order to avoid the congestion caused by commuters or shoppers.

The fact that HGVs are seen to make up a significantly larger proportion of the flow during the low-density overnight periods supports the assertion that the drop in speed at low densities as seen in Figure 4.1 is a result of overnight HGV activity.

Figure 4.6 shows the correlation between HGV percentages and speed. Note that the data have been separated into two sets, “low density” and “high density”. These correspond to hourly observations with density of less than and more than 15 veh/km/lane, respectively. The low-density region shows typically high speeds, dropping gradually as the HGV percentage increases. The high-density region contains many

---

**Figure 4.4** LM299 normalised speed–frequency distribution coloured by day-type groupings. (High-density data (>15 veh/km/lane) are excluded. Gaps correspond to zero counts.)

**Figure 4.5** HGV percentages from MIDAS data for LM299 plotted for two weeks
lower speeds and no HGV percentages above 20%. A simple linear relationship is a good fit for the low-density data.

Figure 4.7 shows the same plot as Figure 4.6, but repeated with data for motorway link LM241 (junctions 46–47 of the M1) – a far less congested stretch of motorway. There are no data at all in the high-density series, but the low-density data show a very similar pattern to those observed for LM299. This linear behaviour suggests that free-flow or low-density speed can be estimated simply by a relationship of the form:

\[ u_f = Hu_h + (1 - H)u_l \]

(Equation 4.1)

where \( H \) is the HGV proportion and \( u_h \) and \( u_l \) are the desired speeds of HGVs and non-HGVs, respectively. The data available for LM299 and LM241 suggest that the desired speed of HGVs is typically \( \sim 85 \text{km/h} \), and that of non-HGVs \( \sim 112 \text{km/h} \).

The implication of these findings is that, at low densities, the interaction between vehicles is minimal and that all vehicles are able to travel at their desired speed. Therefore, the average speed on the link is simply the average of the desired speeds of all the vehicles on that link, as expressed in Equation 4.1. In higher-density traffic, the interactions between vehicles appear to limit the speed to such an extent that the composition of the traffic becomes irrelevant and the HGV percentage shows no correlation with speed. It is possible that the HGV percentage has more subtle effects on traffic flow, such as reducing capacity or inducing congestion at lower flows. The tendency of HGV drivers to avoid times of high demand may restrict the level of available data on such situations, and much more detailed investigations would be required.
5 Modelling the speed–flow–density relationship

5.1 Connections between speed and density

The speed–density relationship \( (u = k^{-1}) \) is the most conceptually accessible of the three possible pairings of speed, flow and density, because it can be interpreted physically in terms of driver response to headway distance between vehicles (\( k \) being the inverse of vehicle spacing). It is typically observed that \( u(k) \) is monotonic and decreasing. This corresponds to the proposition that as the density of vehicles increases, drivers tend to reduce their speed. This is not necessary, however. Vehicles widely enough spaced can travel at their own speed regardless of other traffic (subject to road geometry, vehicle performance and speed limits). Hypothetically, vehicles coupled electronically could increase in density without losing speed, up to a point. If drivers indulge in flocking or racing behaviour then average speed could increase as traffic density increases, though only up to a point. Average speed may also vary because traffic composition is not constant. For example, HGVs tend to form a greater proportion of traffic late at night, so might be expected to lower the average speed, but such effects are not fundamental to traffic behaviour and so should be eliminated from analysis.

As flow increases, vehicles must eventually begin to get in each other’s way, especially if they vary in performance or desired speed. As a result of this interaction, slower vehicles tend to impede quicker vehicles, but the converse, that quicker vehicles tend to encourage slower vehicles to accelerate, is either not true or very much weaker, as well as being discretionary rather than obligatory for safety.

A profusion of macroscopic models for the speed–density relationship have appeared in published literature over the last 75 years (for example, Greenshields, 1935; Greenberg, 1959; Underwood, 1961). Some of these are discussed in detail in the following sections, where it is shown that many of them form a family of solutions to a general differential equation for the reaction of a following vehicle to changes in the speed of its leader (an example of a “stimulus-response” model).

Microscopic car-following models attempt to replicate the behaviour of vehicles in a traffic stream to predict macroscopic traffic dynamics. A common form is that of Gipps (1981), which calculates the maximum safe speed allowing for the maximum deceleration in an emergency. There are many other examples of microscopic models based on car-following algorithms, as cited by Gerlough and Huber (1975), Klar and Wegener (1999), Helbing et al. (2002), Transportation Research Board (2003) and Zhang and Kim (2005).

In reality, there is variation between drivers in their aggressiveness and their responsiveness to their surroundings. In an attempt to circumvent the resulting complexity, Daganzo et al. (1999) introduced a model where there are just two populations, “slugs”, who are relaxed, and “rabbits”, who are in a hurry. While models such as this can be applied and tested in detail on small sections of road, they involve too many behavioural parameters to be calibrated on a whole network of links. If all vehicles are assumed to behave identically, then spacing simply becomes a function of speed. Even so, there are still a large number of parameters that must be specified, and there could be other equally important factors that
might influence behaviour (e.g. road geometry, lane changing, weather). The practical advantage of simple models like those of Greenshields and Greenberg is that all their parameters are linked to quantities that can be measured in aggregation.

A different macroscopic approach is adopted by fluid dynamic models based on the principle of conservation of vehicles, regardless of how $u$ and $k$ change; see, for example, Gerlough and Huber (1975), Cattner et al. (1992), Klar and Wegener (1999), Daganzo et al. (1999), Helbing et al. (2002) and Transportation Research Board (2003). If there is no source or sink of vehicles between two measuring stations at positions along the road $x$ and $x + \Delta x$, then the density of traffic between the two at time $t$ is:

$$k(t) = \frac{N}{\Delta x}$$

(Equation 5.1)

where $N$ is the number of vehicles on the section. After an elapsed time $\Delta t$, the density is equal to:

$$k(t + \Delta t) = q(x + \Delta t) - q(x + \Delta x)\Delta t + N$$

(Equation 5.2)

where $q(x)$ and $q(x + \Delta x)$ are the flows measured at the two stations, respectively. This can be rearranged to give:

$$\Delta k = k(t + \Delta t) - k(t) = \frac{q(x) - q(x + \Delta x)\Delta t}{\Delta x}$$

(Equation 5.3)

Thus, the condition for conservation of vehicles is expressed by Equation 5.4 (left). This is replaced by a partial differential continuity equation in the infinitesimal limit (right):

$$\frac{\Delta q}{\Delta x} + \frac{\Delta k}{\Delta t} = 0 \rightarrow \left( \text{limit} \right) \frac{\partial q}{\partial x} + \frac{\partial k}{\partial t} = 0$$

(Equation 5.4)

However, due to their inherently continuous nature, these models can have difficulty reproducing the behaviour of discrete traffic.

### 5.2 Stimulus-response car-following models

The car-following models of Gazis et al. (1961) have the general form in which the acceleration of a following vehicle reacts to the difference in speed with the vehicle in front in general terms, so that:

- increasing positive speed difference leads to increasing positive acceleration, and vice versa,
- decreasing spacing leads to increasing deceleration, and vice versa, and

- acceleration may also depend directly on the speed of the vehicle.

If vehicle $n$ is ahead of vehicle $n+1$, then these models can be written, following Gazis et al. (1959, 1961) and Easa (1982), as:

$$\ddot{x}_{n+1}(t + \tau) = a \left( \frac{\dot{x}_n(t) - \dot{x}_{n+1}(t)}{\dot{x}_n(t) - \dot{x}_{n+1}(t)} \right)^2$$

(Equation 5.5)

where $l$ and $m$ are parameters, $a_{nm}$ is a dimensioned constant specific to the formulation, $t$ is current time, $T$ is a time step for projection into the future (usually equal or close to an assumed reaction time), $x$ is position and $u$ is speed. They belong to the category of stimulus-response models where the response (change of speed) is driven by the stimulus of the observed position and movement of the vehicle ahead.

Equation 5.5 has the advantage that it can be integrated to yield a family of macroscopic models, along the lines of Equation 2.2. All macroscopic models satisfy the “fundamental relationship” of Equation 2.1, which is:

$$q = uk$$

(Equation 5.6)

A further constraint can be applied from the observation that actual speed-flow relationships are two-valued, so that maximum flow is achieved at some speed, above and below which flow falls. This can be seen as having two causes. At low densities, the flow is low because there is little traffic, regardless of its speed. At high densities, where traffic is forced to adhere to low speeds, flow is low because vehicles pass the measuring point very slowly, and if they stop then the measured flow drops to zero. Any theoretical function that exhibits this behaviour must meet the condition $dq/dk = 0$ for some finite $(q, k)$.

This can be expressed in terms of speed and density by substituting from Equation 5.6 to give:

$$\frac{du}{dk} = \frac{u}{k}$$

(Equation 5.7)

Another important quantity is the wave speed $u_w$, or the speed at which disturbances propagate. For a step change in traffic state, Equation 5.8 (also 2.3) – where subscripts $dn$ and $up$ represent the downstream and upstream states, respectively – must be true in order that vehicles are conserved:

$$u_w = \frac{q_{dn} - q_{up}}{k_{dn} - k_{up}}$$

(Equation 5.8)
Along the $q$–$k$ curve, this is replaced by the continuous form:

$$u_w = \frac{dq}{dk} = u + k \frac{du}{dk}$$

(Equation 5.9)

Addison and Heydecker (2008) point out that only functions for which $d^2q/dk^2$ is positive, ie the $q$–$k$ curve is concave, can give rise to stable shockwaves.\(^\dagger\)

5.3 Specific macroscopic forms of the Gazis, Herman and Rothery models

Integrating Equation 5.4 with various choices of parameters, and applying boundary conditions in the steady state, yields a range of macroscopic models, each characterised by a particular relationship between two of those variables, usually speed and density. Easa (1982) defines a set of five feasible models, out of a possible nine combinations, depending on the choice of integer exponents $l$ and $m$. These models can be arranged into a sequence. While Easa (1982) considered the use of non-integer values of parameters to satisfy boundary conditions (ie measured properties), it is instructive to identify the models that result from integer parameters. Intriguingly, the first model turns out to be the equivalent of the simple linear speed–flow relationship often applied to uncongested traffic (see also discussion earlier in Section 3.6). All the other models emerge as speed–density relationships, which appear to be most appropriate to congested traffic.

5.3.1 Linear speed–flow model ($l = 2, m = 2$)

$$u = u_r - aq$$

(Equation 5.10)

where $u_r$ is the free speed and the slope $a$ is the same as the coefficient in Equation 5.5. Easa (1982) points out that this relationship need never meet the $u = 0$ axis at feasible values of density.

5.3.2 Underwood’s (1961) exponential model ($l = 1, m = 2$)

$$u = u_r \exp(-k/k_o)$$

(Equation 5.11)

where $u_r$ is the “reference” speed and $k_o$ is the “optimal” density. The optimal density represents the point at which flow is maximised according to Equation 5.7. However, for typical motorway conditions $u_r$ needs to be around 300 km/h, much greater than free-flow speed, so does not represent any obvious physical measure (the unrealism of this quantity does not have serious consequences because that part of the function is not used).

5.3.3 Greenshields’ (1935) linear model ($l = 0, m = 2$)

$$u = u_r \left[1 - \frac{k}{k_o}\right]$$

(Equation 5.12)

where $u_r$ is the highest speed supposedly achieved, at zero density, and $k_o$ is the jam density at which speed is zero. This is the simplest model that gives rise to a parabolic speed–flow relationship, necessarily resulting in the speed at maximum flow being $u_r/2$. No real speed–flow relationship is close to this symmetrical shape, especially on multi-lane roads. This difficulty prompted the adoption of “two-regime” models, with a separate model for uncongested traffic, but we have seen that the simplest form of uncongested model arises naturally in the Gazis, Herman and Rothery (GHR) family (Gazis et al., 1961).

5.3.4 Greenberg’s (1959) logarithmic model ($l = 0, m = 1$)

$$u = u_r \ln\left(\frac{k}{k_o}\right)$$

(Equation 5.13)

where $u_r$ is “optimum speed” and $k_o$ is jam density. Assuming that the capacity of a three-lane motorway is somewhere around 6500 veh/h, and the speed at which maximum flow is achieved is around 110 km/h, we need to specify jam density around 160 veh/km, which is less than half real jam density. Increasing the value of $k_o$ greatly increases the maximum flow, to nearly 15 000 veh/h, which is well in excess of what is observed. In fact, plots of $u$ against $\ln(k)$ become non-linear at low speeds, so Greenberg’s model does not explain the observed speed–density relationship. Greenberg’s model also predicts unfeasibly high speeds at densities lower than that associated with $u_r$ (tending to infinity as density tends to zero) so cannot be considered to apply at these densities.

5.3.5 Duncan’s (1979) model ($l = 0, m = 0$)

$$u = q^\left\{\frac{1}{k} - \frac{1}{k_o}\right\}$$

(Equation 5.14)

where $q_o$ is “reference” flow and $k_o$ is jam density. This model arises by spacing vehicles just far enough apart to avoid collision in emergency deceleration allowing for finite reaction time, assuming equal constant deceleration. It can be seen as a highly simplified expression of Gipps’ ideas. Its formulation in Equation 5.5 is unusual because acceleration depends only on speed difference and not on spacing at all. This makes it unlikely to describe traffic moving at higher speeds, where spacing is an important cue. The parameters are not strictly independent, because the ratio $q_o/k_o$ is the upstream speed...
of shockwaves in a traffic jam, which is observed almost universally to be around 20 km/h. This is the only model in the family to give a realistic value for shockwave speed. The speed–flow relationship for a three-lane motorway with nominal capacity around 6500 veh/h is as shown in Figure 5.1 below. The curve is monotonic because Equation 5.7 is satisfied only at \( u = \infty \).

5.4 Comparative properties of the models

5.4.1 Form of speed–density relationships

In speed–density data, the density values tend to change more rapidly as speed decreases, and to spread out at the lowest speeds. For this reason, speed–log density relationships are sometimes plotted. These have the effect of linearising the plots, which are shown for three of the models in Figure 5.1. The red-boxed regions delimit the range of actual data. While the curves are very different, it is hard to decide which best matches real data, although Underwood probably matches best in terms of both linearity at lower densities and the shape of the “tail” at higher densities.

5.4.2 Shockwaves

Under jam conditions, upstream-propagating shockwaves can occur. These can legitimately be called shockwaves, rather than just “change waves”, because they act as boundaries to an extreme condition, where traffic is forced to a halt. At such a boundary, Equation 5.8 for the speed of the boundary as a function of the flows and densities becomes:

\[
    u_s = \frac{-q}{k_j - k}
\]

(Equation 5.15)

where \( q \) and \( k \) are the flow and density where the traffic is moving. Substituting the speed–density relationships of the three models with \( q = uk \), we find that the shockwave speeds they predict are approximately as follows:

- Underwood: \( 0 \)
- Greenshields: \( -u_r \)
- Greenberg: \( -u_o \)
- Duncan: \( -q/o_k \)

As pointed out earlier, of these simpler models only Duncan’s is able to predict anything near the speed of around –19 km/h commonly observed.

5.4.3 Recasting the models in terms of physically meaningful constants

In order to better compare these models, and for them to be of use in modelling, it is useful to replace supposed physical parameters by ones that are actually measured. For example, in Duncan’s model the jam density \( k_j \) can be taken at face value because it leads to realistic shockwave behaviour, while the jam density \( k_j \) in Greenberg’s model need not bear any relation to actual jam density, because the model is determined by conditions around maximum flow. Similarly, the “reference” speeds that turn up in some models have no physical meaning, so are better replaced by measurable values such as optimum speed. On this basis, where in all cases \( q_o = u_o k_o \), the models can be rewritten.

The linear model written in speed–density form is:

\[
    u = \frac{u_r}{1 + au_j k}
\]

(Equation 5.16)

Here, \( u_r \) is the free-flow speed, as in Section 5.3.1. The constant \( a \) has the dimensions of time per vehicle. Its inverse is actually the limiting flow at infinite density, which is a somewhat theoretical concept.

Underwood’s model can be rewritten as:

\[
    u = u_o \exp \left( \frac{1}{k_j} \right)
\]

(Equation 5.17)

thereby eliminating the unphysical parameter \( u_r \), which is in fact identical to \( e^* u_o \).

Greenshields’ model can be rewritten (subject to \( k < 2k_o \)) as:

\[
    u = u_o \left( 2 - \frac{k}{k_o} \right)
\]

(Equation 5.18)
The original $k_j$ in Greenshields’ model is not identified with an independent jam density because it is in fact identical to $2 \cdot k_o$.

Greenberg’s model can be rewritten as:

$$u = u_o \left[ 1 - \ln \left( \frac{k_j}{k_o} \right) \right]$$

(Equation 5.19)

The original $k_j$ in Greenberg’s model is not identified with an independent jam density because it is in fact identical to $e \cdot k_o$.

Duncan’s model can be rewritten as:

$$u = u_o \left[ 1 - \frac{k_j}{k} \right]$$

(Equation 5.20)

The parameter $k_j$ is actual jam density, which can be measured, and $u_j = -q/k_j$ is the shockwave speed (negative meaning upstream). The original parameter $q_j$ is linked to driver response time – it is the inverse of the common response time required to avoid collisions if everyone brakes at the same rate. This cannot be measured so easily but can be calibrated from the shockwave speed. Alternative parameters are $\lambda = 1/k$, representing the space reserved by a vehicle, and $T = 1/q$, representing nominal response time, whereby $u_j = \lambda/T$.

5.4.4 Similarity of Underwood’s, Greenshields’ and Greenberg’s models near “optimum” point

It is interesting to note that the models of Underwood, Greenshields and Greenberg have very similar (and to second order) symmetrical forms when $k = k_o$. Taylor expansion gives:

Underwood:  
$$u = u_o (2 - k/k_o) + u_o \frac{1}{2} (1 - k/k_o)^2 + ...$$

(Equation 5.21)

Greenshields:  
$$u = u_o (2 - k/k_o)$$

(Equation 5.22)

Greenberg:  
$$u = u_o (2 - k/k_o) - u_o \frac{1}{2} (1 - k/k_o)^2 + ...$$

(Equation 5.23)

As a result, there may be little to choose between them at the lower congested densities. Greenberg’s model can be inverted as:

$$\frac{1}{k} = \frac{1}{k_j} \exp \left( \frac{u}{u_o} \right) = \frac{1}{k_j} \left[ 1 + \frac{u}{u_o} + \frac{1}{2} \left( \frac{u}{u_o} \right)^2 + ... \right]$$

(Equation 5.22)

This might suggest an allowance of space for each vehicle that is at least the inverse of jam density, but which also has a reaction-safety component increasing with speed, and a dynamic-safety component increasing with energy, as well as higher-order terms whose significance is not clear. However, the higher-order terms cannot be neglected at speeds near free flow. At the lowest speeds, where we can neglect the second-order and above term, the formula does not reduce to Duncan’s, but differs by a substantial factor, around $k/k_o$, and therefore will not predict correct shockwave speeds.

5.4.5 Seeking a unified approach

Despite stemming from the same root, the five functions discussed appear to divide naturally between coverage of uncongested traffic, traffic around capacity and at lower congested densities, and traffic around jam conditions. Each function is unsuitable or less suitable for other regimes, either because it does not depend critically on parameters associated with those regimes or because it is structurally unable to represent traffic in those regimes. Although it may be possible to establish “mean” values of the microscopic parameters or vary them across the range of conditions, as proposed by Easa (1982), it seems worthwhile to try to combine the basic macroscopic functions, which are individually simple and convenient to work with, in such a way that they predict realistic behaviour over the whole range of speeds and densities.
6 Modelling congested traffic using queuing theory

6.1 Deterministic queuing analysis

The purpose of modelling applied to motorways is to describe or predict what happens over an extended length of road, which can be difficult to observe, as discussed earlier. When modelling congestion, it is usual to consider a single bottleneck – a point at which capacity is restricted compared with the upstream section. Bottlenecks usually occur at particular locations on the road, such as a merge, diverge, lane-drop, roadworks or an incident (although it is not necessary to restrict the definition to static locations).

A bottleneck is said to become "active" when the demand exceeds the capacity of the bottleneck; when this occurs there is bound to be some congestion and delay, for which there are various approaches to modelling. Bottleneck analysis is widely discussed in many published articles, including Gerlough and Huber (1975), Papacostas (1987), Chin (1996a, 1996b) and Gartner et al. (1992).

The simplest models for bottlenecks produce a queue at the active bottleneck, which contains a number of vehicles equal to the difference between demand (Q) and capacity (C) multiplied by the time for which the bottleneck has been active. In these models, the queue just consists of the number of standing vehicles waiting for the capacity to become available for them to pass the bottleneck, referred to as a "vertical queue". This model is satisfactory as long as the queue is relieved by falling demand. If the queue is relieved by removal of the bottleneck, eg by clearing of an incident, then the head of the queue will retreat upstream, resulting in a different clearance time. A further effect is that of physical queue length. As the queue builds, its tail moves upstream to intercept approaching traffic, resulting in a quicker growth rate than would otherwise be the case. These general cases are analysed by "horizontal queuing", which takes account of the speed of approaching vehicles.

Chin (1996a, 1996b) compared vertical and horizontal approaches. In the latter, a region of dense, slow-moving traffic is modelled, the back of which propagates upstream while the bottleneck is active. Chin called this boundary a shockwave, although this is not to be confused with the self-sustaining shockwaves discussed later in this report, and is better called a change wave. After the demand drops below capacity, the back of the queue propagates downstream until it reaches the bottleneck, at which time the queue is dispersed and the bottleneck deactivates. A schematic diagram of this process is shown in Figure 6.1, and a more detailed discussion can be found in May (1990). Where the bottleneck itself is removed, or even where the bottleneck is in motion, the geometry of queue development is more complicated; see, for example, Taylor 2006a and 2006b.

A deterministic queue model ignores the effect of random variations in demand and capacity. Since such variations lead to an equilibrium queue whose size depends on the ratio Q/C only, this is particularly appropriate for motorways where flows are high and the loss of capacity tends to be substantial. Traffic (assumed initially free flowing) approaching the bottleneck is forced to reduce its flow to the capacity of the bottleneck. The only ways this can happen, consistent with the speed–flow relationship, are: (a) by increasing free-flow speed; or (b) by moving to a much lower speed state on the lower half of the relationship. Not only is increasing free-flow speed intuitively impossible, it can be shown that the associated change wave would propagate rapidly downstream, wiping out the region of higher speed – ie it is unstable. The only stable outcome is that a region of slower-moving, higher-density traffic is formed. If demand continues to exceed capacity, more and more vehicles are forced to slow and the queue grows. If demand falls, flow within the queue is maintained at capacity until the queue is dispersed. Queuing traffic is usually assumed to be homogeneous for the sake of simplicity, though this need not be the case. What can be said is that although the causes of the queuing have effects at long range, the interactions that determine its local behaviour are predominantly short range, making it possible to apply the speed–flow models that have been discussed earlier in this report.

Delay can be calculated in various ways, but the delay to each vehicle is essentially the time it takes to move through the queue at reduced speed, less the time it would have taken to travel the same distance at free-flow speed. Overall delay depends primarily on the excess of demand over capacity.

A shockwave can be defined as one where some variable is forced to an extremal value, as for example in stop-start traffic. This is clearly not the case in normal queuing.

![Figure 6.1 Formation and dissipation of a queue consisting of dense, slow-moving traffic upstream of a bottleneck, relaxed by a drop in demand](diagram.png)
and only secondarily on the exact processes in the queue. This approach is very useful and practical; it provides a simple, physical way of obtaining delay and flow from demand and capacity. For this reason, such models are adopted widely for assessing the impact of bottlenecks.

### 6.2 Traffic conservation between homogeneous regions

In a homogeneous region of traffic of known flow and density, such as either half of Figure 6.2, where all variables are the same regardless of where they are measured, the speed of traffic obeys the “fundamental” Equation 2.1, hence:

\[ u = \frac{q}{k} \]

(Equation 6.1)

Therefore, if \( N \) is the number of vehicles on the link, the journey time for a region of length \( l \) is:

\[ JT = \frac{l}{u} = \frac{kl}{q} \cdot \frac{N}{q} \]

(Equation 6.2)

If two such homogeneous regions with different characteristics are placed consecutively, as in Figure 6.2, then as traffic flows from one to the other the boundary between the two regions will move at a speed \( S \) (given by Equation 6.3), which is defined in the same frame of reference as \( u \) and is positive downstream (Chin 1996a, 1996b). This is in effect the macroscopic version of the continuity–conservation equation (Equation 5.4), hence:

\[ S = \frac{q_1 - q_2}{k_1 - k_2} \]

(Equation 6.3)

Note that if \( q_1 = q_2 \), the boundary between the two regions is stationary but the journey time for the combined regions is still given by Equation 6.2, meaning that the region does not actually have to be homogeneous in density or speed. As long as it is homogeneous in flow, the journey time will be conserved. However, if one considers regions of differing flow, the densities become important in determining the propagation of the boundary.

### 6.3 A simple queue model

A simple way of modelling a queue (as used in QUADRO$^\text{a}$) is to assume that a region of dense traffic (the queue) exists immediately upstream of a bottleneck (vertical queue). In this region, the flow is equal to the capacity of the bottleneck and a notional density is chosen to reflect the observed or expected density of slow-moving traffic (Figure 6.3); the speed of this region is then determined using Equation 6.1. The region upstream of the queue is specified as having a speed and flow corresponding to the demand on the route. In this vertical model, it is assumed that the flow into the queue is equal to the demand flow, giving a speed of propagation of the back of the queue of:

\[ S = -\frac{Q - C}{k_2} \]

(Equation 6.4)

Note that, comparing Equations 6.4 and 6.3, it has effectively been assumed that the density in the region upstream of the queue is zero. This is clearly not consistent with the stream equations as finite, non-zero speed and flow values have already been defined. It is possible to demonstrate that, because of this inconsistency, vehicles are not conserved in this model. It is interesting to note that, in this model, the density of the queue will have no effect on journey time, as the number of vehicles joining the queue per unit time is simply equal to \( Q - C \) and the journey time is given by Equation 6.2. However, the delay is affected, as a lower density stretches the queue over a longer distance, which would take longer to traverse under normal conditions, meaning the difference in journey time caused by the queue is reduced.

---

$^a$ QUADRO is a DfT-sponsored computer program maintained and distributed by TRL. Its primary use is in rural areas. It estimates the effects of roadworks in terms of time, vehicle-operating and accident costs on the users of the road. See www.trlsoftware.co.uk
6.4 A more realistic queue model allowing for finite densities

It is very straightforward to modify the model described above so that it obeys the stream equations in both regions and vehicles are conserved (Figure 6.4). The only complication introduced is the fact that the number of vehicles joining the queue per unit time is no longer simply equal to the demand. The model is specified in exactly the same way as above, but instead of assuming a rate of queue growth, as in the previous example, the density of the free-flow region is calculated using the stream equation and the speed of propagation of the back of the queue is given by Equation 6.3. The number of vehicles joining the queue per unit time is therefore given by Equation 6.5:

\[ k_2 S = \frac{Q - C}{k_1 - k_2} k_2 \]

(Equation 6.5)

Unlike the simplified model, this is dependent not only on demand and capacity, but also on the densities of the two regions. Although the journey time at any particular moment is still dependent only on the number of vehicles in each region and the flow, the evolution of the number of vehicles in each region over time is now dependent on \( k_1 \) and \( k_2 \). It is no longer sufficient to simply specify these values because they are determined by the speed–flow–density relationships applying on either side of the change wave.

6.5 Constraining region properties with a speed–flow–density relationship

The model described above assumes the flows of the two regions to be equal to demand and capacity, respectively. This is a perfectly reasonable assumption as the flow through the bottleneck cannot (by definition) exceed capacity, and the flow far from the bottleneck is defined as the demand for the purposes of these models. However, the model also requires a specified free-flow (input) speed and queue density. These variables can have a significant effect on the way the regions evolve, so are important to the model.

In reality, speed and density are seen to be closely related (see Section 3). If it is accepted that there is a deterministic relationship between these two quantities, this acts as a constraint on the properties of the two regions. Given the stream equation (Equation 2.1), and a second equation describing the relationship between any two of the three variables (such as those described in Section 5), it is only necessary to specify one of speed, flow and density to determine the other two. This removes the need to specify an input speed or queue density, as these are now functions of the flow. There are often two possible combinations of speed and density that satisfy the constraints for a given flow. This is illustrated in Figure 6.5, which shows a typical speed–flow relationship, with the points on this curve being consistent with flows equal to demand or capacity.

In the two-region model, the approach can be in state 2 or 3, while the queue can be in either state 1 or 4. In practice, one can suppose the arriving traffic to be free flowing, so it is only in state 2. On the queue side, state 1 represents a higher speed but lower flow than either state 2 or 3, meaning the density of the queue would be lower than that of the approach. Not only is this counter-intuitive, it is also unsustainable, as the boundary between the two regions would have a positive \( S \) according to Equation 6.3, meaning it would propagate downstream until it reached the site of the bottleneck and the region would simply disappear (as discussed in Section 6.1). Therefore, the queue region must be in state 4. In this region, density increases while flow reduces, resulting in an upstream moving change wave that sustains the queue and maintains the bottleneck flow at capacity.
6.6 Downstream of a bottleneck

The traffic downstream of the bottleneck can also be modelled as a homogenous region. In the absence of further bottlenecks or any junctions, the flow must remain equal to that immediately before the bottleneck (also ensuring that the boundary between these two regions remains fixed at the bottleneck).

Referring again to Figure 6.5, this suggests that the traffic can be in either state 1 or state 4 as it leaves the bottleneck. As there is (and can be) no change in flow, the traffic could in principle remain in the congested state 4. However, there is nothing to stop it from changing to state 1 except vehicles ahead. What is observed is that vehicles accelerate, moving vertically on the diagram. If they do this in turn, with a short time lag between them, then they necessarily move apart and their density decreases as their speed increases, in accordance with Equation 2.1. After a time they reach their “target” speed, where the constraints represented by the upper phase of the speed–flow relationship come into effect.

The downstream region would actually experience a negative delay, with speeds being higher than they would have been in the absence of the bottleneck, although to the benefit of a reduced number of vehicles. If this section were long enough, the time saving from higher speed attained might in principle outweigh the delay caused by queuing. It is a real effect. For example, it has been shown that by holding back a portion of traffic during a peak, overall delay can be reduced (system optimum), though only at the expense of increased delay suffered by some individuals (not user optimum). In practice, mixing with other traffic will tend to limit this effect, and if the upper phase of the speed–flow relationship is quite “flat” the benefit will be small anyway, so not worth modelling.

6.7 An example queuing model based on a speed–flow relationship

Section 6 has introduced the concept of a two-region deterministic queuing model with traffic characteristics constrained by a speed–flow relationship. To illustrate this further, an example is provided of how this might be applied to a lane-closure model. In this model, the inputs required are a speed–flow relationship for a single lane, the demand flow and the number of lanes available.

Figures 6.6–6.8 show a three-lane road with a typical single-lane speed–flow relationship with a maximum flow of 2000 veh/h/lane. The figures demonstrate the effect on flow of closing one or two lanes. Note that this model assumes that vehicles are distributed evenly across the available lanes and then treats each lane identically and independently. When one or more lanes is closed, the flow on the remaining lanes is increased accordingly. In the case that the remaining lanes are capable of supporting this flow (as in Figure 6.7), it simply has the effect of moving the traffic into a higher-flow, lower-speed regime. In this regime, the difference in journey time due to the lane closure is simply the difference between the times taken to traverse the closed section at the original and modified levels of flow per lane.

If lane closure results in demand on the remaining lanes or lanes exceeding the maximum flow supported by the speed–flow relationship, vehicles will be forced to wait to enter the remaining lanes at the maximum flow possible. This gives rise to a queue, which can be dealt with exactly as described earlier in Section 6. This situation is illustrated in Figure 6.8. Note that, in order to create a sustainable, high-density queue, the lower-speed solution of the speed–flow relationship must be used in this case.

If one were to use a single-value speed–flow relationship such as those used in COBA, it would not be possible to model the effect of queuing in this way. A queue forms when traffic flow reduces approaching a bottleneck. With a single-value, monotonic, decreasing speed–flow relationship, a reduction in flow must bring about an increase in speed and a reduction in density. It was shown earlier in this section that a queue of lower density than the approaching traffic is unsustainable in reality. Therefore, if one wishes to use a COBA-style single-value speed–flow relationship, it is necessary to define a special queuing state that is not compatible with the speed–flow relationship (as in the examples given in Sections 6.3 and 6.4). The need to define such a state could be said to negate any benefits of the single-value speed–flow relationship.

A model of this kind has the advantage that it does not require the user to specify a capacity for the works site; the model is purely based on the number of lanes available. However, this convenience comes at the price of a significant assumption: that a lane functions identically (and independently of other lanes) regardless of whether it is part of a works site or not. It is quite possible that the speed–flow relationship for a single lane in works may be quite different from a single lane on the same road in non-works conditions. In fact, this is almost certainly the case, as most motorway roadworks will often be accompanied by a 50 mph speed limit. Whether this would simply have the effect of truncating the speed–flow relationship at 50 mph is unclear, and very few data are available to investigate such circumstances. Other effects, such as behavioural changes due to the proximity of a workforce or the possible results of narrowing lanes, are also neglected by the lane-closure model.

There is a second issue that would have to be addressed if a model of this kind was to be implemented: that of sensitivity. Deterministic queuing models are fundamentally dependent on the difference between demand and capacity. The resultant journey time can be quite sensitive to these values and (as discussed in Section 3) these quantities are not necessarily straightforward to measure or define. An additional complication is introduced if capacity is to be defined as the maximum of a speed–flow relationship, as the high-flow region of the speed–flow data typically exhibits significant scatter, meaning the fit in this region is likely to be of low precision. It is therefore likely that the final estimate of journey time would also be of poor precision.

These practical difficulties in successfully calibrating such a model are one reason why single-value speed–flow relationships with separate “queuing” states are found in most current traffic flow models (such as QUADRO). Another reason is that, over a finite period such as a peak, average delay does rise monotonically with overall demand, so for many modelling problems a two-regime approach is not necessary.
Figure 6.6 Evenly distributed flow on three running lanes

Figure 6.7 Single-lane closure causing redistribution of traffic

Figure 6.8 Two-lane closure causing queuing
6.8 Case study: modelling the impact of abnormal loads

Perhaps the ultimate challenge for deterministic queuing apart from network modelling is to describe realistically the effect of a moving bottleneck. There is relatively extensive literature on this problem, but the particular interest at TRL has been in modelling the effect of abnormal loads as they move along a route in varying traffic conditions (see Taylor et al., 2006; Taylor 2006a, 2006b, 2008). This requires consideration of several factors in addition to the usual inputs of traffic demand and capacity:

- The speed of the moving bottleneck.
- The physical extension of the queue (horizontal queuing).
- The traffic state in the queue.

Figure 6.9 shows the MTV plot of a section of the M6 on which a slow-moving abnormal load caused a large queue. Bandwidth limitations at the time meant that data were not available from many of the MIDAS detectors installed on the section (blue bands). However, where MIDAS data were available (narrow bands with dark and light parts), they reveal the region of queuing whose overall shape is triangular. (That a better picture than this is not to hand is evidence that this kind of event is quite rare.)

The shape of the queuing region in space-time is what is expected. The left side of the triangle marks the path of the load, occupying two out of three lanes with its escort, and moving at around 12 mph. The lower right side of the triangle marks the tail of the queue, the boundary between the free-flowing traffic approaching and the traffic backed up behind the load, which propagates in accordance with Equation 6.3 (or 2.3). In this case, the propagation is downstream, which cannot happen in the case of stationary bottlenecks. The top right side of the triangle marks the traffic discharging from the queue after the load has left the motorway. This is also a phase boundary as described by Equation 6.3, but here the wave speed is close to that of shockwaves in queues. Note that the queue discharges from the head rather than shrinking from the tail, because in this case there has been an increase in capacity rather than a decrease in demand.

The traffic state in the model was assumed to be described by Duncan’s relationship (Equation 5.20). It was not possible to measure the traffic state within the queue from the limited data, but important measurements like the demand and the passing flow were measured “live” in this case, enabling the realism of the modelling to be confirmed (more detail is given in Taylor et al., 2006).

This is a particularly simple case because the traffic conditions remain homogeneous throughout. In practice, abnormal loads travel along links of varying standard, at varying speeds in the presence of varying traffic demands. If queues spill back over one or more links upstream of that on which the load is currently, the geometry of the queuing region becomes complicated. In the simple case where queuing forms a homogeneous triangular region, the queue size is described by a single equation, which is the generalisation of the static relationship queue = (demand – capacity) x time:

\[ L = \frac{q_a \left(1 - \frac{u}{u_L}ight) - \mu_c \left(1 - \frac{u}{u_d}ight)}{1 - \frac{K}{K_0}} \]

(Equation 6.6)

Here, the effective demand \(q\) is modified by the speed of the load \(u_L\) in relation to the speed of the approaching traffic \(u_a\), and the effective capacity \(\mu_c\) is also modified by the speed of the load in relation to the speed of receding traffic.
traffic \(u_d\), which could be slightly different from \(u_a\). The effect is to reduce the queue size. However, the denominator, which depends on the relationship between the densities of approaching and queuing traffic, tends to increase the queue size. The physical extent of the queue is derived by dividing the queue size in vehicles by the traffic density:

\[
X = \frac{L}{k_v}
\]  
(Equation 6.7)

In practice, the calculation is usually done the other way around, obtaining the extent of the queue by geometric means. This also simplifies the calculation of delay, which for each individual vehicle depends on the average excess time it spends queuing. This in turn can be calculated from the average extent of the queue and the speeds of approaching and queuing traffic, respectively. In this case, \(X\) is the maximum queue, so the average is half (where the flows, speeds and densities are related by Equation 6.1):

\[
d = \frac{X}{2} \left( \frac{1}{d_v} - \frac{1}{d_a} \right)
\]  
(Equation 6.8)

An important point is that if the states of approaching and queuing traffic are similar, it is possible to have a very extended queue, which causes very little delay. Horizontal queue modelling generally predicts greater delay than vertical queue modelling, but in the case of the moving bottleneck the concept of a vertical static queue has no meaning. Although it is possible to devise “adjusted” demand and capacity to put into a conventional queue model, the non-intuitive nature of such adjustments and the difficulty of making equivalent measurements mean it is safer to do the full calculation. A critical part of the calculation is determining the speed and density in the different traffic phases, which depend on the type of understanding covered earlier in this report.

### 7 Conclusions

#### 7.1 Summary

In recent years, the UK has seen a shift in emphasis away from road building and towards the use of technology to fully utilise the capacity of existing roads. In order to develop technologies that effectively optimise traffic flow, it is vital to understand the underlying processes that determine the speed and flow of traffic.

The relationship between the macroscopic properties of traffic – speed, flow and density – is not trivial. The relationship arises from the compounded behaviours of individual drivers responding to their immediate environment. The understanding of these relationships in traffic may be about to be accelerated by the availability of large amounts of data from networks of automatic detectors.

It has been shown in this report that the data from these detectors support the idea (as implied by the “classical” models of Underwood, Greenberg, Greenshields and others) that average speed decreases with increasing traffic density, although the data from HATRIS suggest that this decrease is only significant above a certain density. Below this, speed is independent of density. This is not a feature of any of the “classical” models, suggesting that none of them fully describes traffic over the full range of possible densities. However, it has been shown that all of them can be derived from the GHR car-following equation, as can a simple linear model where speed is independent of density.

The data are also seen to support the concept of a two-value speed–flow curve where a given flow can be sustained at two different speeds, converging at a single optimum speed and flow. High-resolution MIDAS data support the theory that these two speeds correspond to congested and free-flowing states. The processes governing the transition between these states in unobstructed traffic are not fully understood. However, the transitions in the presence of a bottleneck are consistent with deterministic queuing theory.

The individual measurements of speed and flow stored in HATRIS show considerable scatter – with a large range of speeds observed for any given density. It has been shown that some of this scatter can be attributed to the dependence of the relationships themselves on external influences such as traffic composition. If traffic flow is to be modelled accurately, it may be necessary to determine how such external influences affect the parameters of the relationships.

Present understanding of traffic flow theory has already allowed road operators to better understand the behaviour of road users and take practical steps to improve performance. For example, the impact of abnormal loads on the UK motorways network has been modelled and quantified, allowing the HA to make better-informed decisions on the movement of such loads. The continued deployment of technologies such as Controlled Motorways and Active Traffic Management mean that the value of fully understanding the mechanics of traffic flow is becoming greater by the day.
7.2 Areas for future investigation
While the increase in monitoring technology has led to improvements in modelling and the development of tools such as MTV, there is still considerable scope for further research into the fundamental processes that drive traffic flow. The huge amounts of data now available present a previously unavailable opportunity to thoroughly test and refine theoretical models. Models based on individual driver behaviour have proved difficult to calibrate because they depend on internal variables, which are not directly measurable in aggregate traffic. Macroscopic models, by and large, can be calibrated from aggregate measurements, so in this regard may be more useful, but still need to be shown to have merit as physical explanations, not just as convenient formulations. The work described in this Insight Report is beginning to tease out the properties of different models, and to identify what needs to be measured and what essential features models should possess. These are key points that must be addressed if a model is to be useful and meaningful across all applications. The matter will not be settled by either theoretical model formulation or data fitting alone, but will need the convergence of both.

Acknowledgments
The work described in this Insight Report was carried out in the Network Performance Group of the Transport Research Laboratory. The authors are grateful to Dr Alan Stevens, who carried out the technical review and auditing of this report. Much of the work described in this report was carried out under a project (020(387)HTRL – Operational Support and Development of Network Tools) funded by the Highways Agency South East Regional Intelligence Unit, for Dr George Skrobianski. The authors would like to acknowledge the valuable insights and contributions of Dr Skrobianski, and of colleagues at TRL who worked on that project: Dr Shanta Tuladhar, Tim Rees and Wenqun Wang. Motorway Traffic Viewer (MTV) was developed by Peter Still and Dr Brian Williams and subsequently by the MTV team at TRL.
References


Acikelik R (1991, revised 2000). Travel time functions for transportation purposes; Davidson’s function, its time-dependent form and an alternative travel time function. Australian Road Research, 21 (3) 49–59.


Further reading


It seems intuitively obvious that the more traffic that tries to use a given section of road, the slower it must move, but the precise mechanisms behind this relationship are surprisingly elusive. The availability of large amounts of data collected automatically raises the possibility of validating theories and models. This Insight Report examines the features of some actual data and speed–flow–density relationships, and “classical” models of speed, flow and density in the context of the wealth of detailed traffic data now available. Data from detectors support the idea that average speed decreases with increasing traffic density, although the data suggest that this decrease is only significant above a certain density. Below this, speed is virtually independent of density. This is not a feature of “classical” macroscopic models, suggesting that none of them fully describes traffic over the full range of possible densities. The role of speed, flow and density in queuing theory is also examined, including a case study of modelling a moving bottleneck.

Other titles in this series

INS001 Improving the stability of slopes using a spaced piling technique. D R Carder. 2009