TRAFFIC STREAM CHARACTERISTICS
BY FRED L. HALL[4]

[4]Professor, McMaster University, Department of Civil Engineering and Department of Geography, 1280 Main Street West, Hamilton, Ontario, Canada L8S 4L7.

CHAPTER 2 - Frequently used Symbols

k	density of a traffic stream in a specified length of road
L	length of vehicles of uniform length
c_k	constant of proportionality between occupancy and density, under certain simplifying assumptions
k_i	the (average) density of vehicles in substream *I*
q_i	the average rate of flow of vehicles in substream *I*
$\bar{u}$	average speed of a set of vehicles
A	$A(x,t)$ the cumulative vehicle arrival function over space and time
k_j	jam density, i.e. the density when traffic is so heavy that it is at a complete standstill
u_f	free-flow speed, i.e. the speed when there are no constraints placed on a driver by other vehicles on the road

2.
TRAFFIC STREAM CHARACTERISTICS

Author's note: This material has benefited greatly from the assistance of Michael Cassidy, of the University of California at Berkeley. He declined the offer to be listed as a co-author of the chapter, although that would certainly have been warranted. With his permission, and that of the publisher, several segments of this material have been reproduced directly from his chapter, "Traffic Flow and Capacity", which appears as Chapter 6 in ***Handbook of transportation science****, edited by Randolph W. Hall, and published by Kluwer Academic Publishers in 1999. That material appears in italics below, in section 2.1. The numbering of Figures and equations has been altered from his numbers to correspond to numbering within this chapter.*

In this chapter, properties that describe highway traffic are introduced, such as flow, density, and average vehicle speed; issues surrounding their measurements are discussed; and various models that have been proposed for describing relationships among these properties are presented. Most of the work dealing with these relationships has been concerned with uninterrupted traffic flow, primarily on freeways or expressways, but the general relationships will apply to all kinds of traffic flow.

This chapter starts with a section on definitions of key variables and terms. Because of the importance of measurement capability to theory development, the second section deals with measurement issues, including historical developments in measurement procedures, and criteria for selecting good measurement characteristics. The third section discusses a number of the bivariate models that have been proposed in the past to relate key variables. That is followed by a short section on attempts to deal simultaneously with the three key variables. The final, summary section provides links to a number of the other chapters in this monograph.

2.1 Definitions and terms

This section provides definitions of some properties commonly used to characterize traffic streams, together with some generalized definitions that preserve useful relations among these properties. Before turning to the definitions, however, a useful graphical tool is introduced, namely trajectories plotted on a time-space diagram. This is followed by definitions of traffic stream properties, time-mean and space-mean properties, and the generalized definitions. Following that is a discussion of the relationship between density, which is often used in the generalized definitions, and occupancy, which is measured by many freeway systems. The final topic covered is three-dimensional representations of traffic streams. The discussion in this section follows very closely or repeats some of that in the chapter by Cassidy (1999), which in turn credits notes composed by Newell (unpublished) and a textbook written by Daganzo (1997).

2.1.1 The Time-Space Diagram.

***The Time-Space Diagram**. Objects are commonly constrained to move along a one-dimensional guideway, be it, for example, a highway lane, walkway, conveyor belt, charted course or flight path. Thus, the relevant aspects of their motion can often be described in cartesian coordinates of time, t, and space, x. Figure 2-1 illustrates the trajectories of some objects traversing a facility of length L during time interval T; these objects may be vehicles, pedestrians or cargo. Each trajectory is assigned an integer label in the ascending order that the object would be seen by a stationary observer. If one object overtakes another, their trajectories may exchange labels, as shown for the fourth and fifth trajectories in the figure. Thus, the ℓth trajectory describes the location of a reference point (e.g. the front end) of object ℓ as a function of time t,* $x_\ell(t)$.

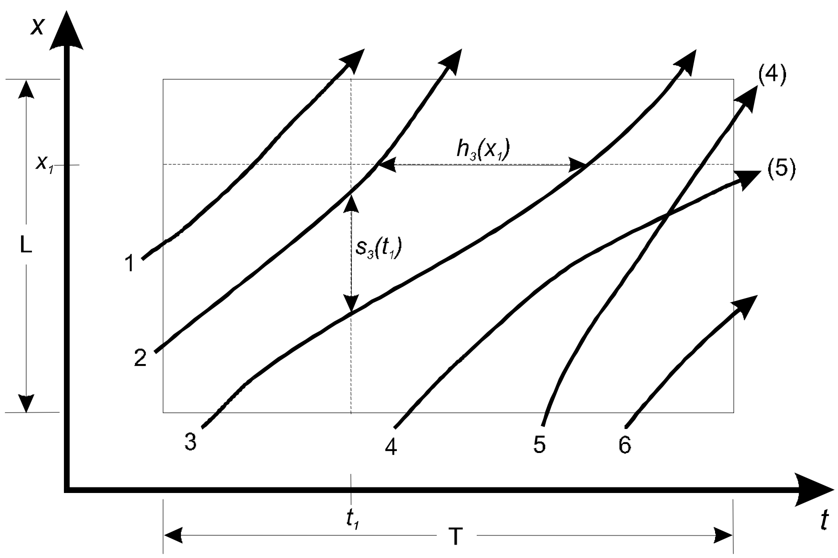

Figure 2-1. Time-space diagram.

The characteristic geometries of trajectories on a time-space diagram describe the motion of objects in detail. These diagrams thus offer the most complete way of displaying the observations that may have actually been measured along a facility. As a practical matter, however, one is not likely to collect all the data needed to construct trajectories. Rather, time-space diagrams derive their (considerable) value by providing a means to highlight the key features of a traffic stream using only coarsely approximated data or hypothetical data from "thought experiments."

2.1.2 Definitions of Some Traffic Stream Properties. *It is evident from Figure 2-1 that the slope of the ℓth trajectory is object ℓ's instantaneous velocity,* $v_\ell(t)$*, i.e.,*

$$v_\ell(t) \equiv dx_\ell(t)/dt \,, \qquad (2.1)$$

and that the curvature is its acceleration. Further, there exist observable properties of a traffic stream that relate to the times that objects pass a fixed location, such as location x_1*, for example.*

These properties are described with trajectories that cross a horizontal line drawn through the time-space diagram at x_1.

Referring to Figure 2-1, the headway of some ith object at x_1, $h_i(x_1)$, *is the difference between the arrival times of i and i-1 at* x_1, *i.e.,*

$$h_i(x_1) \equiv t_i(x_1) - t_{i-1}(x_1) \tag{2.2}$$

Flow at x_1 *is m, the number of objects passing* x_1, *divided by the observation interval T,*

$$q(T, x_1) \equiv m/T. \tag{2.3}$$

For observation intervals containing large m,

$$\sum_{i=1}^{m} h_i(x_1) \approx T \tag{2.4}$$

and thus,

$$q(T, x_1) \approx \frac{1}{\frac{1}{m}\sum_{i=1}^{m} h_i(x_1)} = \frac{1}{\bar{h}(x_1)}, \tag{2.5}$$

i.e., flow is the reciprocal of the average headway.

Analogously, some properties relate to the locations of objects at a fixed time, as observed, for example, from an aerial photograph. These properties may be described with trajectories that cross a vertical line in the t-x plane. For example, the spacing of object j at some time t_1, $s_j(t_1)$, *is the distance separating j from the next downstream object; i.e.,*

$$s_j(t_1) \equiv x_{j-1}(t_1) - x_j(t_1). \tag{2.6}$$

Density at instant t_1 *is n, the number of objects on a facility at that time, divided by L, the facility's physical length; i.e.,*

$$k(L, t_1) \equiv n/L. \tag{2.7}$$

If the L contains large n,

$$\sum_{j=1}^{n} s_j(t_1) \approx L \tag{2.8}$$

and

$$k(L, t_1) \approx \frac{1}{\frac{1}{n}\sum_{j=1}^{n} s_j(t_1)} = \frac{1}{\bar{s}(t_1)}, \tag{2.9}$$

giving a relation between density and the average spacing parallel to that of flow and the average headway.

2.1.3 Time-Mean and Space-Mean Properties. *For an object's attribute α, where α might be its velocity, physical length, number of occupants, etc., one can define an average of the m objects passing some fixed location x_1 over observation interval T,*

$$\alpha(T, x_1) = \frac{1}{m}\sum_{i=1}^{m}\alpha_i(x_1), \qquad (2.10)$$

i.e., a time-mean of attribute α. If α is headway, for example, α(T, x_1) is the average headway or the reciprocal of the flow.

Conversely, the space-mean of attribute α at some time t_1, α(L, t_1), is obtained from the observations taken at that time over a segment of length L, i.e.,

$$\alpha(L, t_1) = \frac{1}{n}\sum_{j=1}^{n}\alpha_j(t_1). \qquad (2.11)$$

If, for example, α is spacing, α(L, t_1) is the average spacing or the reciprocal of the density.

For any attribute α, there is no obvious relation between its time and space means. The reader may confirm this (using the example of α as velocity) by envisioning a rectangular time-space region L × T traversed by vehicles of two classes, fast and slow, which do not interact. For each class, the trajectories are parallel, equidistant and of constant slope; such conditions are said to be ***stationary****.*

The fraction of fast vehicles distributed over L as seen on an aerial photograph taken at some instant within T will be smaller than the fraction of fast vehicles crossing some fixed point along L during the interval T. This is because the fast vehicles spend less time in the region than do the slow ones. Analogously, one might envision a closed loop track and note that a fast vehicle passes a stationary observer more often than does a slow one.

2.1.4 Generalized Definitions of Traffic Stream Properties. *To describe a traffic stream, one usually seeks to measure properties that are not sensitive to the variations in the individual objects (e.g. the vehicles or their operators) without averaging-out features of interest. This is the trade-off inherent in choosing between short and long measurement intervals, as previously noted. It was partly to address this trade-off that Edie (1965, 1974) proposed some generalized definitions of flow and density that averaged these properties in the manner described below.*

To begin this discussion, the thin, horizontal rectangle in Figure 2-2 corresponds to a fixed observation point. As per its conventional definition provided earlier, the flow at this point is m / T, where m = 4 in the figure. Since this point in space is a region of temporal duration T and elemental spatial dimension dx, the flow can be expressed equivalently as $\frac{m \cdot dx}{T \cdot dx}$. The denominator is the euclidean area of the thin horizontal rectangle, expressed in units of distance × time. The numerator

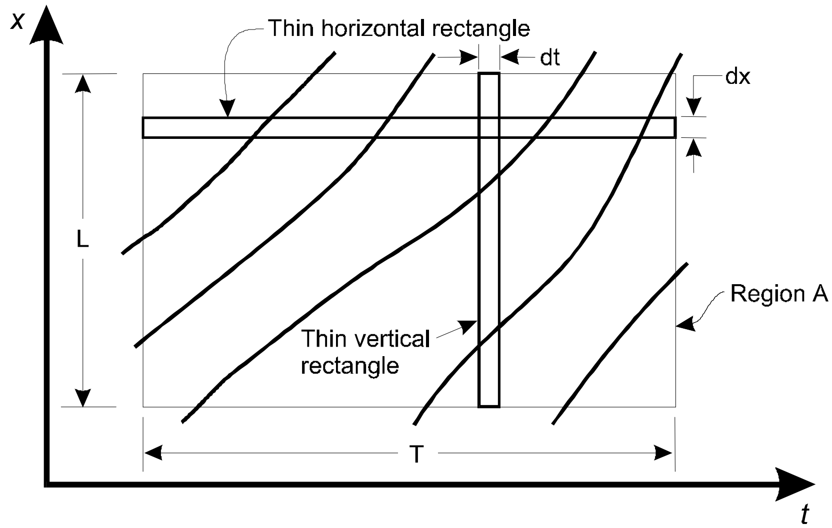

Figure 2-2. Trajectories in time-space region.

is the total distance traveled by all objects in this thin region, since objects cannot enter or exit the region via its elementally small left and right sides.

That flow, then, is the ratio of the distance traveled in a region to the region's area is valid for any time-space region, since all regions are composed of elementary rectangles. Taking, for example, region A in Figure 2-2, Edie's generalized definition of the flow in A, q(A), is d(A) / |A|, where d(A) is the total distance traveled in A and |A| is used to denote the region's area.

As the analogue to this, the thin, vertical rectangle in Figure 2-2 corresponds to an instant in time. As per its conventional definition, density is n / L (where n = 2 in this figure) and this can be expressed equivalently as $\frac{n \cdot dt}{L \cdot dt}$. *It follows that Edie's generalized definition of density in a region A, k(A), is t(A) / |A|, where t(A) is the total time spent in A.*

It should be clear that these generalized definitions merely average the flows collected over all points, and the densities collected at each instant, within the region of interest. Dividing this flow by this density gives d(A) / t(A), which can be taken as the average velocity of objects in A, v(A). The reader will note that, with Edie's definitions, the average velocity is the ratio of flow to density. Traffic measurement devices, such as loop detectors installed beneath the road surface, can be used to measure flows, densities and average vehicle velocities in ways that are consistent with these generalized definitions. Discussion of this is offered in Cassidy and Coifman (1997).

As a final note regarding v(A), when A is taken as a thin horizontal rectangle of spatial dimension dx, the time spent in the region by object i is dx / v_i, *where* v_i *is i's velocity. Thus, for this thin,* *horizontal region A, t(A) =* $dx\sum_{i=1}^{m}\frac{1}{v_i}$. *Given that for the same region, d(A) = m·dx, the generalized mean velocity becomes*

$$v(A)=\frac{d(A)}{t(A)}=\frac{1}{\frac{1}{m}\sum_{i=1}^{m}\frac{1}{v_i}}, \tag{2.12}$$

i.e., the reciprocal of the mean of the reciprocal velocities, or the harmonic mean velocity. The $1/v_i$ *is often referred to as the pace of i,* p_i*, and thus*

$$v(A)=\left[\frac{1}{m}\sum_{i=1}^{m}p_i\right]^{-1}. \tag{2.13}$$

Eq. 2.15 applies for regions with L > dx provided that all i span the L and that each p_i *(or* v_i*) is i's average over the L.*

It follows that when conditions in a region A are stationary, the harmonic mean of the velocities measured at a fixed point in A is the v(A). By the same token, the v(A) is the space-mean velocity measured at any instant in A (provided, again, that conditions are stationary).

2.1.5 The Relation Between Density and Occupancy. *Occupancy is conventionally defined as the percentage of time that vehicles spend atop a loop detector. It is a commonly-used property for describing highway traffic streams; it is used later in this chapter, for example, for diagnosing freeway traffic conditions. In particular, occupancy is a proxy for density. The following discussion demonstrates that the former is merely a dimensionless version of the latter.*

One can readily demonstrate this relation by adopting a generalized definition of occupancy analogous to the definitions proposed by Edie. Such a definition is made evident by illustrating each trajectory with two parallel lines tracing the vehicle's front and rear (as seen by a detector) and this is exemplified in Figure 2-3. The (generalized) occupancy in the region A, $\rho(A)$*, can be taken as the fraction of the region's area covered by the shaded strips in the figure. From this, it follows that the* $\rho(A)$ *and the k(A) are related by an average of the vehicle lengths. This average vehicle length is, by definition, the area of the shaded strips within A divided by the t(A); i.e., it is the ratio of the* $\rho(A)$ *to the k(A),*

$$average\ vehicle\ length=\frac{\rho(A)}{k(A)}=\frac{area\ of\ the\ shaded\ strips}{|A|}\cdot\frac{|A|}{t(A)}. \tag{2.14}$$

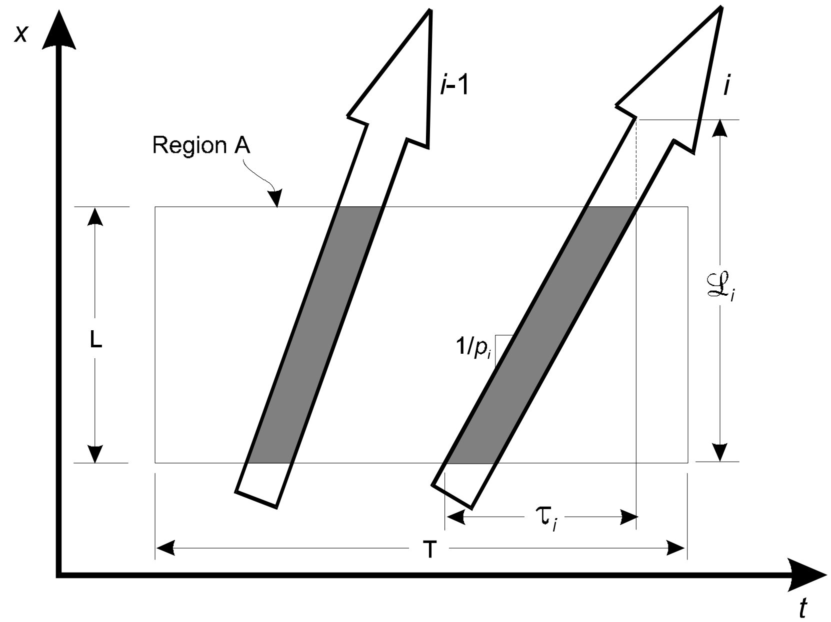

Figure 2-3. Trajectories of vehicle fronts and rears.

Notably, an average of the vehicle lengths also relates the k(*A*) *to* ρ, *where the latter is the occupancy as conventionally defined (i.e., the percentage of time vehicles spend atop the detector). Toward illustrating this relation, the L in Figure 6-4 is assumed to be the length of road "visible" to the loop detector, the so-called detection zone. The T is some interval of time; e.g. the interval over which the detector collects measurements. The time each ith vehicle spends atop the detector is denoted as* τ_i. *Thus, if m vehicles pass the detector during time T, the*

$$\rho = \frac{\sum_{i=1}^{m} \tau_i}{T}.$$

As shown in Figure 6-4, $\mathcal{L}_i$ is *the summed length of the detection zone and the length of vehicle i. Therefore,*

$$\sum_{i=1}^{m} \tau_i = \sum_{i=1}^{m} \mathcal{L}_i \cdot \frac{1}{v_i} = \sum_{i=1}^{m} \mathcal{L}_i p_i \tag{2.17}$$

if the front end of each i has a constant v_i *over the distance* $\mathcal{L}_i$. *Since*

$$\frac{\sum_{i=1}^{m} \tau_i}{T} = \frac{\frac{1}{m} \cdot \sum_{i=1}^{m} \tau_i}{\frac{1}{m} \cdot T} = q(A) \cdot \frac{1}{m} \cdot \sum_{i=1}^{m} \tau_i, \tag{2.16}$$

it follows that

$$\rho = q(A) \cdot \frac{1}{m} \sum_{i=1}^{m} \mathscr{L}_i \cdot p_i ,$$

$$\rho = q(A) \cdot \frac{1}{v(A)} \left[v(A) \cdot \frac{1}{m} \sum_{i=1}^{m} \mathscr{L}_i \cdot p_i \right],$$

$$\rho = k(A) \cdot \left[\frac{\sum_{i=1}^{m} \mathscr{L}_i \cdot p_i}{\sum_{i=1}^{m} p_i} \right], \tag{2.17}$$

where the term in brackets is the average vehicle length relating ρ *to the* $k(A)$; *it is the so-called average effective vehicle length weighted by the paces. If pace and vehicle length are uncorrelated, the term in brackets in (2.17) can be approximated by the unweighted average of the vehicle lengths in the interval T.*

When measurements are taken by two closely spaced detectors, a so-called speed trap, the p_i *are computed from each vehicle's arrival times at the two detectors. The* $\mathscr{L}_i$ *are thus computed by assuming that the* p_i *are constant over the length of the speed trap. When only a single loop detector is available, vehicle velocities are often estimated by using an assumed average value of the (effective) vehicle lengths.*

2.1.6 Three-Dimensional Representation of Vehicle Streams. *It is useful to display flows and densities using a three-dimensional representation described by Makagami et al.* (*1971*). *For this representation, an axis for the cumulative number of objects, N, is added to the t-x coordinate system so that the resulting surface N(t, x) is like a staircase with each trajectory being the edge of a step. As shown in Figure 2-4, curves of cumulative count versus time are obtained by taking cross-sections of this surface at some fixed locations and viewing the exposed regions in the t-N plane. Analogously, cross-sections at fixed times viewed in the N-x plane reveal curves of cumulative count versus space.*

Figure 2-4 shows cumulative curves at two locations and for two instants in time. The former display the trip times of objects and the time-varying accumulations between the two locations, as labeled on the figure. These cumulative curves can be transformed into a queueing diagram (as described in Chapter 5) by translating the curve at upstream x_1 *forward by the free-flow (i.e., the undelayed) trip time from* x_1 *to* x_2. *Also displayed in Figure 2-4, the curves of cumulative count versus space show the number of objects crossing a fixed location during the interval* $t_2 - t_1$ *and the distances traveled by individual objects during this same interval.*

If one is dealing with many objects so that measuring the exact integer numbers is not important, it is advantageous to construct the cumulative curves with piece-wise linear approximations; e.g. the curves may be smoothed using linear interpolations that pass through the crests of the steps. The time-dependent flows past some location are the slopes of the smoothed curve of t versus N *constructed at that location (Moskowitz, 1954; Edie and Foote, 1960; Newell, 1971, 1982). Analogously, the location-dependent densities at some instant are the negative slopes of a smoothed curve of N versus x; densities are the negative slopes because objects are numbered in the reverse direction to their motion.* This three-dimensional model has been applied by Newell (1993) in work on kinematic waves in traffic (see Chapter 5). In addition, Part I of his paper contains some historical notes on the use of this approach for modelling.

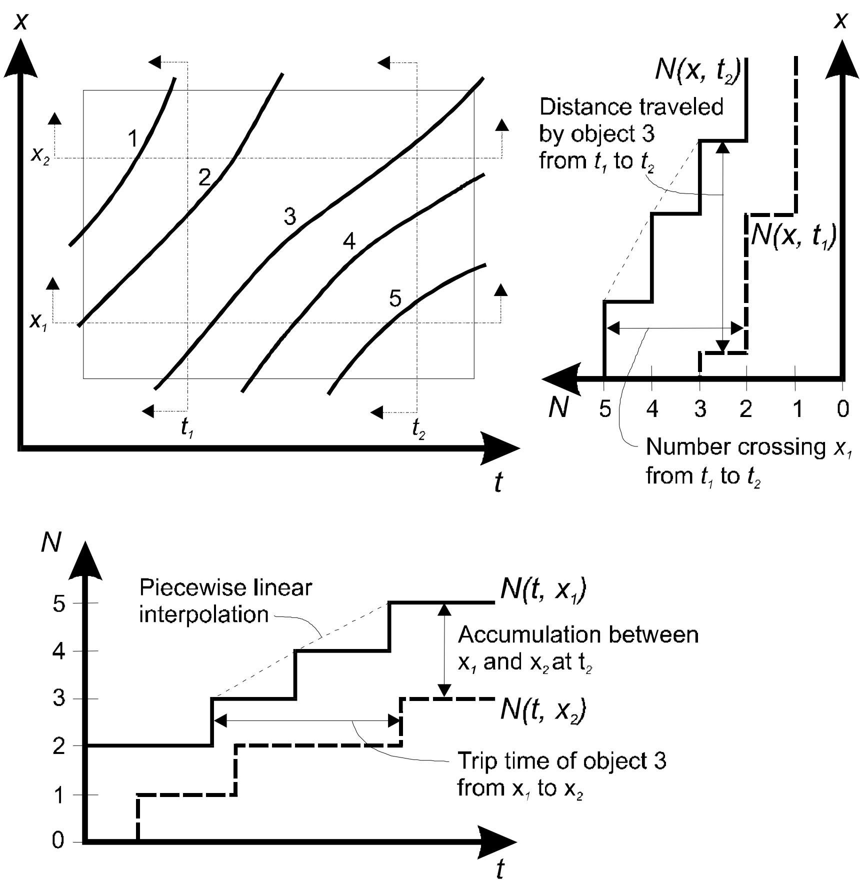

Figure 2-4. Three-dimensional representation.

2.2 Measurement issues

This section addresses issues related to the measurement of the key variables defined in the previous section. Four topics are covered. First, there is a discussion of measurement procedures that have traditionally been used. Second, potential measurement errors arising from the mismatch between the definitions and the measurement methods are discussed. Third, measurement difficulties that can potentially arise from the particular location used for collecting measurements are considered. The final topic is the nature of the time intervals from which to collect data, drawing on the cumulative curves just presented.

2.2.1 Measurement procedures Measurement technology for obtaining traffic data has changed over the 60-year span of interest in traffic flow, but most of the basic procedures remain largely the same. Five measurement procedures are discussed in this section:

- measurement at a point;
- measurement over a short section (by which is meant less than about 10 meters (m);
- measurement over a length of road [usually at least 0.5 kilometers (km)];
- the use of an observer moving in the traffic stream; and
- wide-area samples obtained simultaneously from a number of vehicles, as part of Intelligent Transportation Systems (ITS).

The first two were discussed with reference to the time-space diagram in Fig 2.2, and the remaining three are also readily interpreted on that diagram. Details on each of these methods can be found in the ITE's Manual of Transportation Engineering Studies (Robertson, 1994). The wide-area samples from ITS are similar to having a number of moving observers at various points and times within the system. New developments such as this will undoubtedly change the way some traffic measurements are obtained in the future, but they have not been in operation long enough to have a major effect on the material to be covered in this chapter.

Measurement at a point by hand tallies or pneumatic tubes, was the first procedure used for traffic data collection. This method is easily capable of providing volume counts and therefore flow rates directly, and can provide headways if arrival times are recorded. The technology for making measurements at a point on freeways changed over 30 years ago from using pneumatic tubes placed across the roadway to using point detectors (May et al. 1963; Athol 1965). The most commonly used point detectors are based on inductive loop technology, but other methods in use include microwave, radar, photocells, ultrasonics, and closed circuit television cameras.

Except for the case of a stopped vehicle, speeds at a 'point' can be obtained only by radar or microwave detectors: dx/dt obviously requires some dx, however small. (Radar and microwave frequencies of operation mean that a vehicle needs to move only about one centimeter during the speed measurement.) In the absence of such instruments for a moving vehicle, a second observation location is necessary to obtain speeds, which moves the discussion to that of measurements over a short section.

Measurements over a short section Early studies used a second pneumatic tube, placed very close to the first, to obtain speeds. More recent systems have used paired presence detectors, such as inductive loops spaced perhaps five to six meters apart. With video camera technology, two detector 'lines' placed close together provide the same capability for measuring speeds. All of these presence detectors continue to provide direct measurement of volume and of time headways, as well as of speed when pairs of them are used.

Most point detectors currently used, such as inductive loops or microwave beams, take up space on the road, and are therefore a short section measurement. These detectors also measure occupancy, which was not available from earlier technology. This variable is available because the loop gives a continuous reading (at 50 or 60 Hz usually), which pneumatic tubes or manual counts could not do. Because occupancy depends on the size of the detection zone of the instrument, the measured occupancy may differ from site to site for identical traffic, depending on the nature and construction of the detector.

Measurements along a length of road come either from aerial photography, or from cameras mounted on tall buildings or poles. On the basis of a single frame from such sources, only density can be measured. The single frame gives no sense of time, so neither volumes nor speed can be measured. Once several frames are available, as from a video-camera or from time-lapse photography over short time intervals, speeds and volumes can also be measured, as per the generalized definitions provided above.

Despite considerable improvements in technology, and the presence of closed circuit television on many freeways, there is very little use of measurements taken over a long section at the present time. The one advantage such measurements might provide would be to yield true journey times over a lengthy section of road, but that would require better computer vision algorithms (to match vehicles at both ends of the section) than are currently possible. There have been some efforts toward the objective of collecting journey time data on the basis of the details of the 'signature' of particular vehicles or platoons of vehicles across a series of loops over an extended distance (Kühne and Immes 1993), but few practical implementations as yet.

The moving observer method was used in some early studies, but is not used as the primary data collection technique now because of the prevalence of the other technologies. There are two approaches to the moving observer method. The first is a simple floating car procedure in which speeds and travel times are recorded as a function of time and location along the road. While the intention in this method is that the floating car behaves as an average vehicle within the traffic stream, the method cannot give precise average speed data. It is, however, effective for obtaining qualitative information about freeway operations without the need for elaborate equipment or procedures. One form of this approach uses a second person in the car to record speeds and travel times. A second form uses a modified recording speedometer of the type regularly used in long-distance trucks or buses. One drawback of this approach is that it means there are usually significantly fewer speed observations than volume observations. An example of this kind of problem appears in Morton and Jackson (1992).

The other approach was developed by Wardrop and Charlesworth (1954) for urban traffic measurements and is meant to obtain both speed and volume measurements simultaneously. Although the method is not practical for major urban freeways, it is included here because it may be of some value for rural expressway data collection, where there are no automatic systems. While it is not appropriate as the primary mode of data collection on a busy freeway, there are some useful points that come out of the literature that should be noted by those seeking to obtain average speeds through the moving car method.

The method developed by Wardrop and Charlesworth is based on a survey vehicle that travels in both directions on the road. The formulae allow one to estimate both speeds and flows for one direction of travel. The two formulae are

$$q = \frac{(x + y)}{(t_a + t_w)} \qquad (2.18)$$

$$\bar{t} = t_w - \frac{y}{q} \qquad (2.19)$$

where

q is the estimated flow on the road in the direction of interest,

x is the number of vehicles traveling in the direction of interest, which are met by the survey vehicle while traveling in the opposite direction,

y is the net number of vehicles that overtake the survey vehicle while traveling in the direction of interest (i.e. those passing minus those overtaken),

t_a is the travel time taken for the trip against the stream,

t_w is the travel time for the trip with the stream, and

$\bar{t}$ is the estimate of mean travel time in the direction of interest.

Wright (1973) revisited the theory behind this method. His paper also serves as a review of the papers dealing with the method in the two decades between the original work and his own. He finds that, in general, the method gives biased results, although the degree of bias is not significant in practice, and can be overcome. Wright's proposal is that the driver should fix the journey time in advance, and keep to it. Stops along the way would not matter, so long as the total time taken is as determined prior to travel. Wright's other point is that turning traffic (exiting or entering) can upset the calculations done using this method. This fact means that the route to be used for this method needs to avoid major exits or entrances. It should be noted also that a large number of observations are required for reliable estimation of speeds and flows; without that, the method has very limited precision.

2.2.2 Error caused by the mismatch between definitions and usual measurements

The overview in the previous section described methods that have historically been used to collect observations on the key traffic variables. As was mentioned within that section, the methods do not always accord with the definitions of these variables, as they were presented in Section 2.1. One of the most common methods for collecting these data currently is based on inductive loops embedded in the roadway. When speed data are also to be collected, a pair of closely spaced loops (often called a speed trap) is needed, a known distance apart. Equivalent systems are based on microwave beams that cover a part of a roadway surface. Cassidy and Coifman (1997) point out the criteria that need to be met for the data from these systems to meet the definitional requirements.

In practical terms, these criteria can be reduced to ensuring that any vehicle entering the speed trap also clears it within the time interval for which the data are obtained – or that the number of vehicles in the time interval is quite large. Neither of these criteria is met by the ordinary implementation of inductive loop speed trap detectors. Whether or not the last vehicle entering a speed trap will clear it within the interval is a simple random variable. There is no guarantee that this criterion will regularly be met. If the number of vehicles were very large, this would not be an issue, but because of the need for timely information, many systems poll the loop detector controllers at least every minute, with some systems going to 30 second or even 20 second data collection. For intervals of that size, the volume counts on a single freeway lane will be at most 40 (or 25 or 20) vehicles, which is not large enough to overcome the error introduced by missing one of the vehicle's speeds. While the data from these detector systems are certainly good enough for operational decision-making, the data may give misleading results if used directly with these equations because of the mismatch between the collection and the definitions.

2.2.3 Importance of location to the nature of the data

Almost all of the bivariate models to be discussed represent efforts to explain the behaviour of traffic variables over the full range of operation. In turning the models from abstract representations into numerical models with specific parameter values, an important practical question arises: can one expect that the data collected will cover the full range that the model is intended to cover? If the answer is no, then the difficult question follows of how to do curve fitting (or parameter estimation) when there may be essential data missing.

This issue can be explained more easily with an example. At the risk of oversimplifying a relationship prior to a more detailed discussion of it, consider the simple representation of the speed-flow curve as shown in Figure 2.5, for three distinct sections of roadway. The underlying curve is assumed to be the same at all three locations. Between locations A and B, a major entrance ramp adds considerable traffic to the road. If location B reaches capacity due to this entrance ramp volume, there will be a queue of traffic on the mainstream at location A. These vehicles can be considered to be waiting their turn to be served by the bottleneck section immediately downstream of the entrance ramp. The data superimposed on graph A reflect the situation whereby traffic at A had not reached capacity before the added ramp flow caused the backup. There is a good range of uncongested data (on the top part of the curve), and congested data concentrated in one area of the lower part of the curve. The flows for that portion reflect the capacity flow at B less the entering ramp flows.

At location B, the full range of uncongested flows is observed, right out to capacity, but the location never becomes congested, in the sense of experiencing stop-and-go traffic. It does, however, experience congestion in the sense that speeds are below those observed in the absence of the upstream congestion. Drivers arrive at the front end of the queue moving very slowly, and accelerate away from that point, increasing speed as they move through the bottleneck section. This segment of the speed-flow curve has been referred to as queue discharge flow, QDF (Hall et al. 1992). The particular speed observed at B will depend on how far it is from the front end of the queue (Persaud and Hurdle 1988). Consequently, the only data that will be observed at B are on the top portion of the curve, and at some particular speed in the QDF segment.

If the exit ramp between B and C removes a significant portion of the traffic that was observed at B, flows at C will not reach the levels they did at B. If there is no downstream situation similar to that between A and B, then C will not experience congested operations, and the data observable there will be as shown in Figure 2.5.

None of these locations taken alone can provide the data to identify the full speed-flow curve. Location C can help to identify the uncongested portion, but cannot deal with capacity, or with congestion. Location B can provide information on the uncongested portion and on capacity operation, but cannot contribute to the discussion of congested operations. Location A can provide some information on both uncongested and congested operations, but cannot tell anything about capacity operations. This would all seem obvious enough. A similar discussion appears in Drake et al (1967). It is also explained by May (1990). Other aspects of the effect of location on data patterns are discussed by Hsu and Banks (1993). Yet a number of important efforts to fit data to theory have ignored this key point (for example Ceder and May 1976; Easa and May 1980). They have recognized that location-A data are needed to fit the congested portion of the curve, but have not recognized that at such a location data are missing that are needed to identify capacity. Consequently, discussion of bivariate models will look at the nature of the data used in each study, and at where the data were collected (with respect to bottlenecks) in order to evaluate the theoretical ideas. It is possible that the apparent need for several different models, or for different parameters for the same model at different locations, or even for discontinuous models instead of continuous ones, arose because of the nature (location) of the data each was using.

2.2.4 Selecting intervals from which to extract data

In addition to the location for the data, there are also several issues relating to the time intervals for collecting data. The first is the issue of obtaining representative data. By examining trends on the cumulative curves, one can observe how flows and densities change with time and space. By selecting flows and densities as they appear on the cumulative curves, their values may be taken over intervals that exhibit near-constant slopes. In this way, the values assigned to these properties are not affected by some arbitrarily selected measurement interval(s). Choosing intervals arbitrarily is undesirable because data extracted over short measurement intervals are highly susceptible to the effects of statistical fluctuations while the use of longer intervals may average-out the features of interest. Further discussion and demonstration of this in the context of freeway traffic is offered in (Cassidy, 1998).

There is also an issue of how many observations are needed to obtain good estimates of key variables such as the capacity. *A bottleneck's capacity, q_{max}, is the maximum flow it can sustain for a very long time (in the absence of any influences from restrictions further downstream). It can be expressed mathematically as*

$$q_{\max} \equiv \lim_{T\to\infty}\left(\frac{N_{\max}}{T}\right)^{\wedge}, \qquad (2.20)$$

where N_{max} denotes that the vehicles counted during very long time T discharged through the bottleneck at a maximum rate. The engineer assigns a capacity to a bottleneck by obtaining a value for the estimator q_{max} (since one cannot actually observe a maximum flow for a time period approaching infinity). It is desirable that the expected value of this estimator equal the capacity, $E(q_{max})\tilde{}\ q_{max}$. For this reason, one would collect samples (i.e., counts) immediately downstream of an active bottleneck so as to measure vehicles discharging at a maximum rate. The amount q_{max} can *deviate from q_{max} is controlled by the sample size, N. A formula for determining N to estimate a bottleneck's capacity to a specified precision is derived below.*

To begin this derivation, the estimator may be taken as

$$\hat{q}_{\max} = \sum_{m=1}^{M} n_m / (M \cdot \tau), \tag{2.21}$$

where n_m is the count collected in the mth interval and each of these M intervals has a duration of τ. If the $\{n_m\}$ can be taken as independent, identically distributed random variables (e.g. the counts were collected from consecutive intervals with τ sufficiently large), then the variance of $\hat{q}_{max}$ can be expressed as

$$variance(\hat{q}_{\max}) = \frac{1}{\tau^2}\left[\frac{variance(n)}{M}\right] = \frac{1}{T}\left[\frac{variance(n)}{\tau}\right] \tag{2.22}$$

since q_{max} is a linear function of the independent n_m and the (finite) observation period T is the denominator in (2.21).

The bracketed term variance(n) / τ is a constant. Thus, by multiplying the top and bottom of this quotient by E(n), the expected value of the counts, and by noting that E(n) / τ = q_{max}, one obtains

$$\text{variance}(\hat{q}_{\max}) = \frac{\gamma}{T}, \tag{2.23}$$

where γ is the index of dispersion; i.e., the ratio variance(n)/E(n).

The variance($\hat{q}_{max}$) is the square of the standard error. Thus, by isolating T in (2.23) and then multiplying both sides of the resulting expression by q_{max}, one arrives at

$$q_{\max} \cdot T = \frac{\gamma}{\varepsilon^2}, \tag{2.24}$$

where $q_{max} \cdot T \approx N$, the number of observations (i.e., the count) needed to estimate capacity to a specified percent error ε. Note, for example, that ε = 0.05 to obtain an estimate within 5 percent of q_{max}. The value of γ may be estimated by collecting a presample and, notably, N increases rapidly as ε diminishes.

The expression $N = \gamma/\varepsilon^2$ may be used to determine an adequate sample size when vehicles, or any objects, discharging through an active bottleneck exhibit a nearly stationary flow; i.e., when the cumulative count curve exhibits a nearly constant slope. If necessary, the N samples may be obtained by concatenating observations from multiple days. Naturally, one would take samples during time periods thought to be representative of the conditions of interest. For example, one should probably not use vehicle counts taken in inclement weather to estimate the capacity for fair weather conditions.

2.3 Bivariate models

This section provides an overview of work to establish relationships among pairs of the variables described in the opening section. Some of these efforts begin with mathematical models; others are primarily empirical, with little or no attempt to generalize. Both are important components for understanding the relationships. For reasons discussed above, two aspects of these efforts are emphasized here: the measurement methods used to obtain the data; and the location at which the measurements were obtained.

2.3.1 Speed-flow models

The speed-flow relationship is the bivariate relationship on which there has been the greatest amount of work recently, so is the first one discussed. This section starts from current understanding, which provides some useful insights for interpreting earlier work, and then moves to a chronological review of some of the major models.

The bulk of then recent empirical work on the relationship between speed and flow (as well as the other relationships) was summarized in a paper by Hall, Hurdle, and Banks (1992). In it, they proposed the model for traffic flow shown in Figure 2.6. This figure is the basis for the background speed-flow curve in Figure 2.5.

It is useful to summarize some of the antecedents of the understanding depicted in Figure 2.6. The initial impetus came from a paper by Persaud and Hurdle (1988), in which they demonstrated (Figure 2.7) that the vertical line for queue discharge flow in Figure 2.6 was a reasonable result of measurements taken at various distances downstream from a queue. This study was an outgrowth of an earlier one by Hurdle and Datta (1983) in which they raised a number of questions about the shape of the speed-flow curve near capacity.

Additional empirical work dealing with the speed-flow relationship was conducted by Banks (1989, 1990), Hall and Hall (1990), Chin and May (1991), Wemple, Morris and May (1991), Agyemang-Duah and Hall (1991) and Ringert and Urbanik (1993). All of these studies supported the idea that speeds remain nearly constant even at quite high flow rates. Another of the important issues they dealt with is one that had been around for over thirty years (Wattleworth 1963): is there a reduction in flow rates within the bottleneck at the time that the queue forms upstream? Figure 2.6 shows such a drop on the basis of two studies. Banks (1991a, 1991b) reports roughly a 3% drop from pre-queue flows, on the basis of nine days of data at one site in California. Agyemang-Duah and Hall (1991) found about a 5% decrease, on the basis of 52 days of data at one site in Ontario. This decrease in flow is often not observable, however, as in many locations high flow rates do not last long enough prior to the onset of congestion to yield the stable flow values that would show the drop.

Two empirical studies in Germany support the upper part of the curve in Figure 2.6 quite well. Heidemann and Hotop (1990) found a piecewise-linear 'polygon' for the upper part of the curve (Figure 2.8). Unfortunately, they did not have data beyond 1700 veh/hr/lane, so could not address what happens at capacity. Stappert and Theis (1990) conducted a major empirical study of speed-flow relationships on various kinds of roads. However, they were interested only in estimating parameters for a specific functional form,

$$V = (A - e^{BQ})\, e^{-c} - K\, e^{dQ} \qquad (2.25)$$

where V is speed, Q is traffic volume, c and d are constant "Krummungs factors" taking values between 0.2 and 0.003, and A, B, and K are parameters of the model. This function tended to

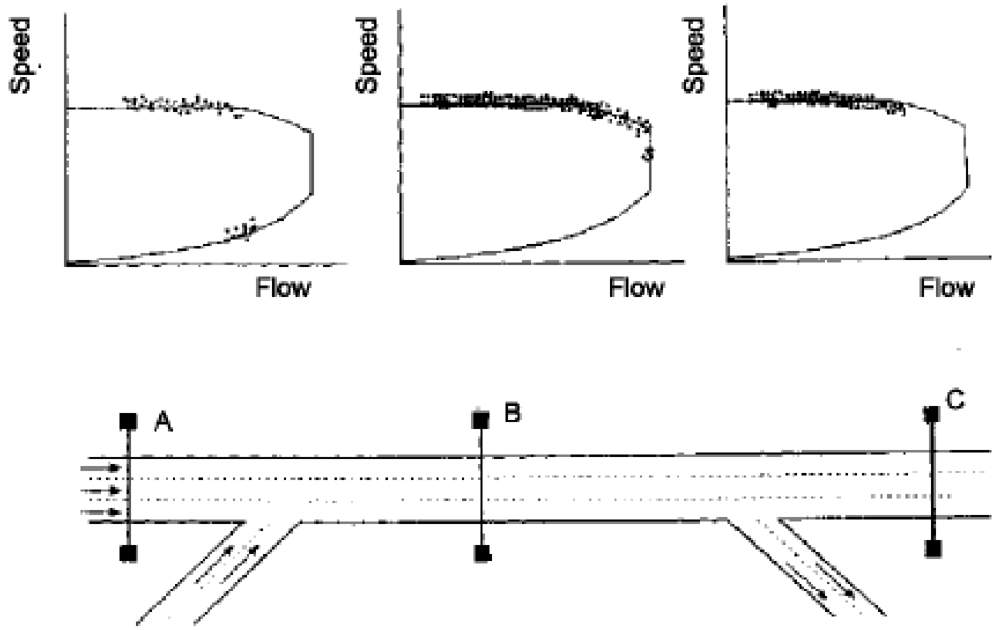

Figure 2. 5
Effect of Measurement Location on Nature of Data
(Similar to figures in May 1990 and Hall et al. 1992).

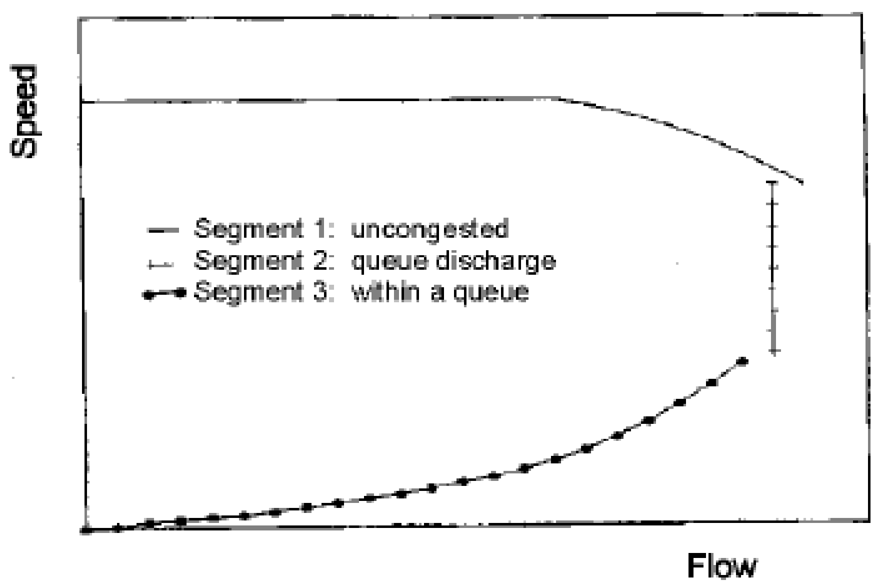

Figure 2. 6
Generalized Shape of Speed-Flow Curve
Proposed by Hall, Hurdle, & Banks
(Hall et al. 1992).

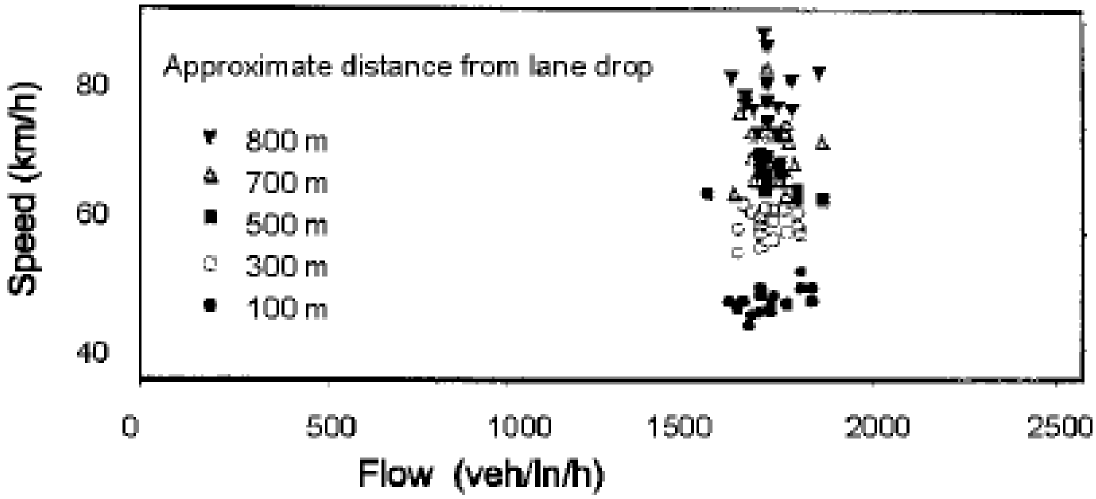

Figure 2. 7
Speed-Flow Data for Queue Discharge Flow at Varied Distances Downstream from the Head of the Queue (Modified from Persaud and Hurdle 1988).

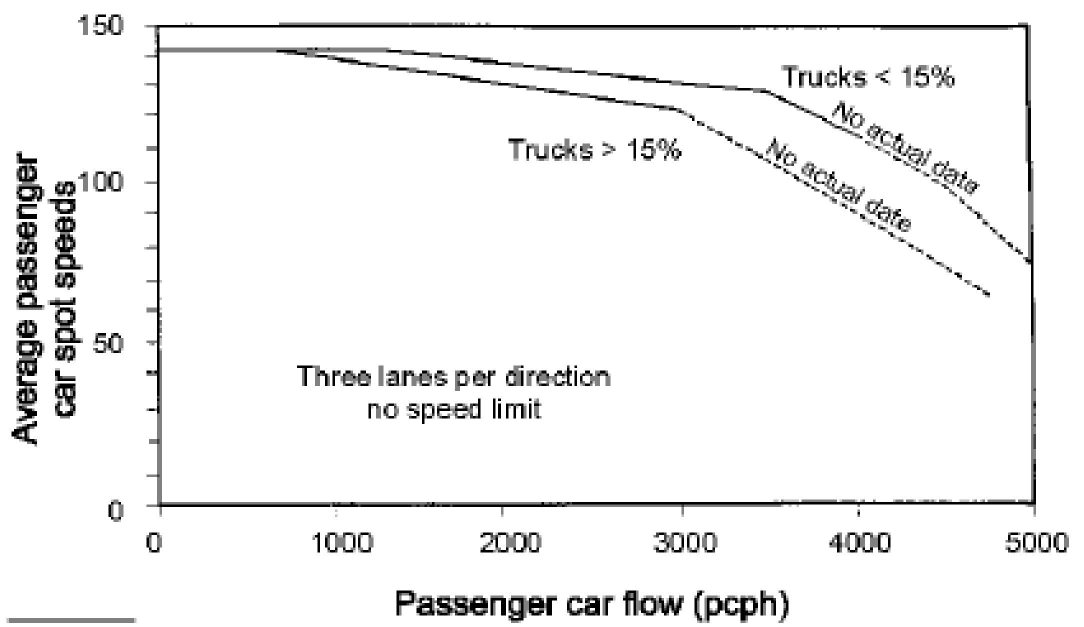

Figure 2. 8
Results from Fitting Polygon Speed-Flow Curves to German Data (Modified and translated from Heidemann and Hotop 1990).

give the kind of result shown in Figure 2.9, despite the fact that the curve does not accord well with the data near capacity. In Figure 2.9, each point represents a full hour of data, and the graph represents five months of hourly data. Note that flows in excess of 2200 veh/h/lane were sustained on several occasions, *over the full hour*.

The problem for traffic flow theory is that these curves are empirically derived. There is not really any theory that would explain these particular shapes, except perhaps for Edie et al.(1980), who propose qualitative flow regimes that relate well to these curves. The task that lies ahead for traffic flow theorists is to develop a consistent set of equations that can replicate this reality. The models that have been proposed to date, and will be discussed in subsequent sections, do not necessarily lead to the kinds of speed-flow curves that data suggest are needed.

It is instructive to review the history of depictions of speed-flow curves in light of this current understanding. Probably the seminal work on this topic was the paper by Greenshields in 1935, in which he derived the following parabolic speed-flow curve on the basis of a linear speed-density relationship together with the equation, flow = speed * density:

$$q = k_j (u - \frac{u^2}{u_f}) \tag{2.26}$$

Figure 2.10 contains that curve, and the data it is based on, redrawn. The numbers against the data points represent the "number of 100-vehicle groups observed", on Labor Day 1934, in one direction on a two-lane two-way road (p. 464). In counting the vehicles on the road, every 10th vehicle started a new group (of 100), so there is 90% overlap between two adjacent groups (p. 451): the groups are not independent observations. Equally important, the data have been grouped in flow ranges of 200 veh/h and the averages of these groups taken prior to plotting. The one congested point, representing 51 (overlapping) groups of 100 observations, came from a different roadway entirely, with different cross-section and pavement, which were collected on a different day.

These details are mentioned here because of the importance to traffic flow theory of Greenshields' work. The parabolic shape he derived was accepted as the proper shape of the curve for decades. In the 1965 Highway Capacity Manual, for example, the shape shown in Figure 2.10 appears exactly, despite the fact that data displayed elsewhere in the 1965 HCM showed that contemporary empirical results did not match the figure. In the 1985 HCM, the same parabolic shape was retained, although broadened considerably. In short, Greenshields' model dominated the field for over 50 years, despite the fact that by current standards of research the method of analysis of the data, with overlapping groups and averaging prior to curve-fitting, would not be acceptable.

There is another criticism that can be addressed to Greenshields' work as well, although it is one of which a number of current researchers seem unaware. Duncan (1976; 1979) has shown that calculating density from speed and flow, fitting a line to the speed-density data, and then converting that line into a speed-flow function, gives a biased result relative to direct estimation

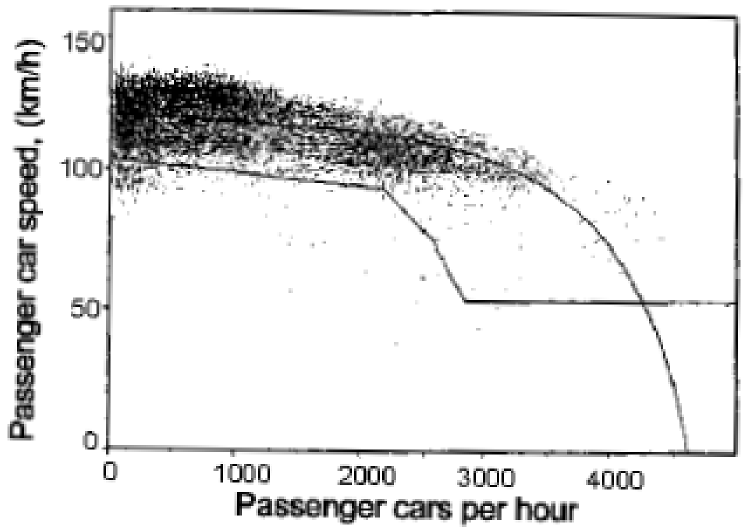

Figure 2. 9
Data for Four-Lane German Autobahns (Two-Lanes per Direction), as reported by Stappert and Theis (1990).

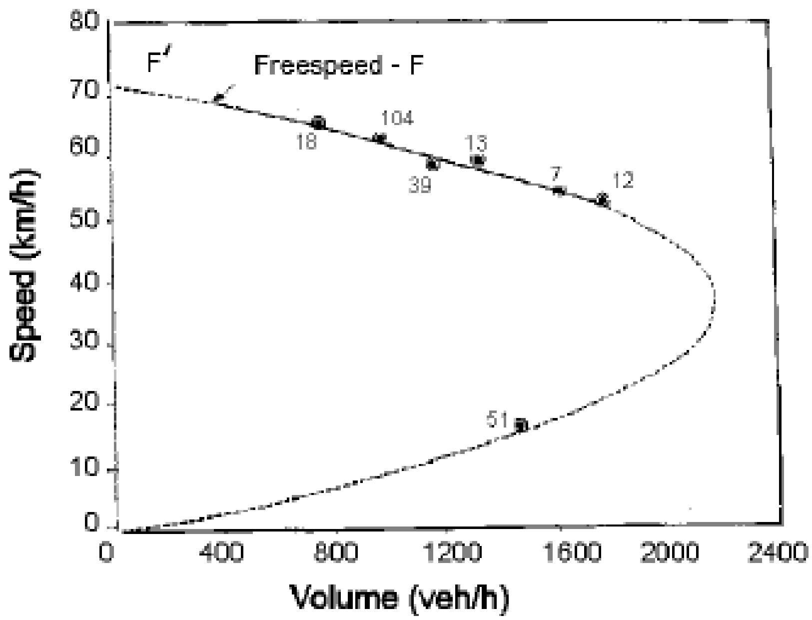

Figure 2.10
Greenshields' Speed-Flow Curve and Data (Greenshields 1935).

of the speed-flow function. This is a consequence of three things discussed earlier: the non-linear transformations involved in both directions, the stochastic nature of the observations, and the inability to match the time and space measurement frames exactly.

It is interesting to contrast the emphasis on speed-flow models in recent years with that 20 years ago, as represented in TRB SR 165 (Gerlough and Huber 1975). In that volume, the discussion of speed-flow models took up less than a page of text, and none of the five accompanying diagrams dealt with freeways. (Three dealt with the artificial situation of a test track.) In contrast, five pages and eleven figures were devoted to the speed-density relationship. Speed-flow models are important for freeway management strategies, and will be of fundamental importance for intelligent vehicle/highway systems (IVHS) implementation of alternate routing; hence there is considerably more work on this topic than on the remaining two bivariate topics. Twenty and more years ago, the other topics were of more interest. As Gerlough and Huber stated the matter (p. 61), "once a speed-concentration model has been determined, a speed-flow model can be determined from it." That is in fact the way most earlier speed-flow work was treated (including that of Greenshields). Hence it is sensible to turn to discussion of speed-concentration models, and to deal with any other speed-flow models as a consequence of speed-concentration work, which is the way they were developed.

2.3.2 Speed-concentration models

Greenshields' (1935) linear model of speed and density was mentioned in the previous section. It is:

$$u = u_f(1 - k/k_j) \tag{2.27}$$

where u_f is the free-flow speed, and k_j is the jam density. The measured data were speeds and flows; density was calculated using equation 2.27. The most interesting aspect of this particular model is that its empirical basis consisted of half a dozen points in one cluster near free-flow speed, and a single observation under congested conditions (Figure 2.11). The linear relationship comes from connecting the cluster with the single point: "since the curve is a straight line it is only necessary to determine accurately two points to fix its direction" (p. 468). What is surprising is not that such simple analytical methods were used in 1935, but that their results (the linear speed-density model) have continued to be so widely accepted for so long. While there have been studies that claimed to have confirmed this model, such as that in Figure 2.12a (Huber 1957), they tended to have similarly sparse portions of the full range of data, usually omitting both the lowest flows, and flow in the range near capacity. There have also been a number of studies that found contradictory evidence, most importantly that by Drake, et al. (1965), which will be discussed in more detail subsequently.

A second early model was that put forward by Greenberg (1959), showing a logarithmic relationship:

$$u = u_m \ln(k/kj) \tag{2.28}$$

The paper showed the fit of the model to two data sets, both of which visually looked very reasonable. However, the first data set was derived from speed and headway data on individual vehicles, which "was then separated into speed classes and the average headway was calculated for each speed class" (p. 83). In other words, the vehicles that appear in one data point (speed class) may not even have been travelling together! While a density can always be calculated as the reciprocal of average headway, when that average is taken over vehicles that may well not have been travelling together, it is not clear what that density is meant to represent. The second data set used by Greenberg was Huber's. This is the same data that appears in Figure 2.12a; Greenberg's graph is shown in Figure 2.12b. Visually, the fit is quite good, but Huber reported an R of 0.97, which does not leave much room for improvement.

A third model from the same period is that by Edie (1961). His model was an attempt to deal with a discontinuity that had consistently been found in data near the critical density, which “suggested the existence of some kind of change of state” (p. 72). He proposed two linear models for the two states of flow. The first related density to the logarithm of velocity above the “optimum velocity”, i.e. “non-congested flow”. The second related velocity to the logarithm of spacing (i.e. the inverse of density) for congested flow. The model was fitted to the same Lincoln Tunnel data as used by Greenberg.

These three forms of the speed-density curve, plus four others, were investigated in an empirical test by Drake et al. in 1967. The test used data from the middle lane of the Eisenhower Expressway in Chicago, one-half mile (800 m) upstream from a bottleneck whose capacity was only slightly less than the capacity of the study site. This location was chosen specifically in order to obtain data over as much of the range of operations as possible. A series of 1224 1-minute observations were initially collected. The measured data consisted of volume, time-mean speed, and occupancy. Density was calculated from volume and time-mean speed. A sample was then taken from among the 1224 data points in order to create a data set that was uniformly distributed along the density axis, as is assumed by regression analysis of speed on density.

The intention in conducting the study was to compare the seven speed-density hypotheses statistically, and thereby to select the best one. In addition to Greenshields' linear form, Greenberg's exponential curve, and Edie’s two-part logarithmic model, the other four investigated were a two-part and three-part piecewise linear models, Underwood's (1961) transposed exponential curve, and a bell-shaped curve. Despite the intention to use "a rigorous structure of falsifiable tests" (p. 75) in this comparison, and the careful way the work was done, the statistical analyses proved inconclusive: "almost all conclusions were based on intuition alone since the statistical tests provided little decision power after all" (p. 76). To assert that intuition alone was the basis is no doubt a bit of an exaggeration. Twenty-one graphs help considerably in differentiating among the seven hypotheses and their consequences for both speed-volume and volume-density graphs.

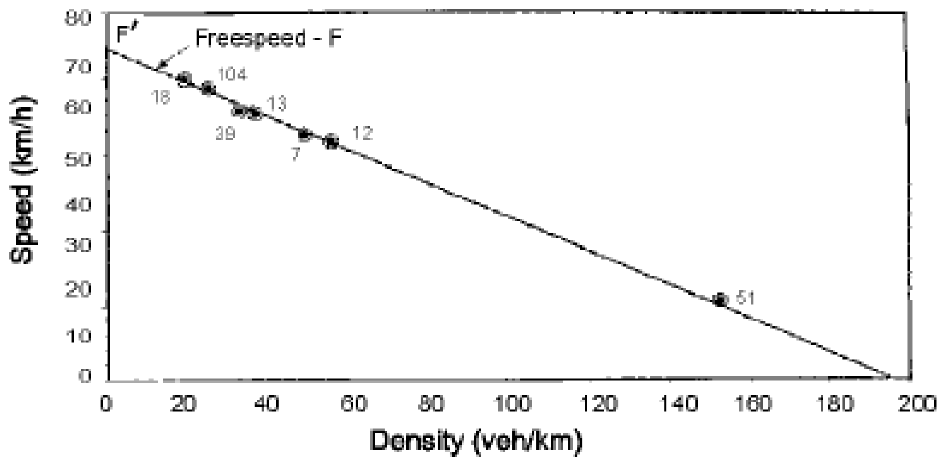

Figure 2.11
Greenshields' Speed-Density Graph and Data (Greenshields 1935).

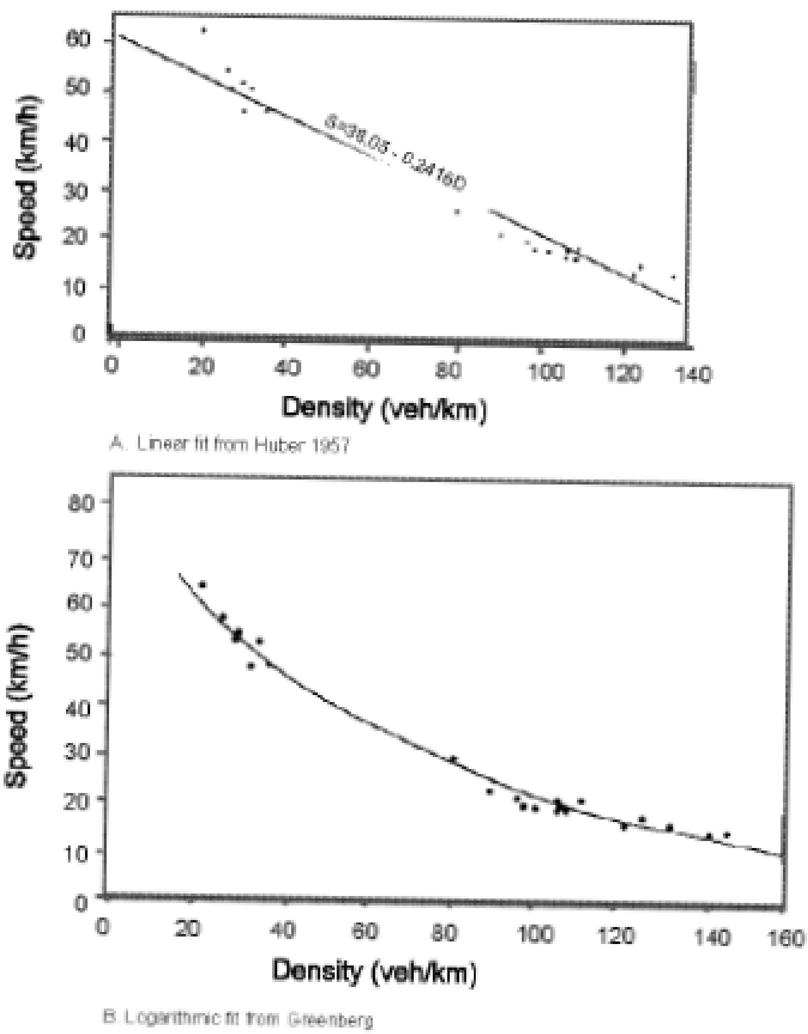

Figure 2.12
Speed-Density Data from Merritt Parkway and Fitted Curves.

Figure 2.13 provides an example of the three types of graphs used, in this case the ones based on the Edie model. Their comments about this model (p. 75) were: "The Edie formulation gave the best estimates of the fundamental parameters. While its R^2 was the second lowest, its standard error was the lowest of all hypotheses." One interesting point with respect to Figure 2.13 is that the Edie model was the only one of the seven to replicate capacity operations closely on the volume-density and speed-volume plots. The other models tended to underestimate the maximum flows, often by a considerable margin, as is illustrated in Figure 2.14, which shows the speed-volume curve resulting from Greenshields' hypothesis of a linear speed-density relationship. (It is interesting to note that the data in these two figures are quite consistent with the currently accepted speed-flow shape identified earlier in Figures 2.5 and 2.6.) The overall conclusion one might draw from the Drake et al. study is that none of the seven models they tested provide a particularly good fit to or explanation of the data, although it should be noted that they did not state their conclusion this way, but rather dealt with each model separately.

There are four additional issues that arise from the Drake et al. study that are worth noting here. The first is the methodological one identified by Duncan (1976; 1979), and discussed earlier with regard to Greenshields' work. Duncan showed that calculating density from speed and flow data, fitting a speed-density function to that data, and then transforming the speed-density function into a speed-flow function results in a curve that does not fit the original speed-flow data particularly well. This is the method used by Drake et al, and certainly most of their resulting speed-flow functions did not fit the original speed-flow data very well. Duncan's 1979 paper expanded on the difficulties to show that minor changes in the speed-density function led to major changes in the speed-flow function, suggesting the need for further caution in using this method of double transformations to get a speed-flow curve.

The second is that of the data collection location, as discussed above. The data were collected upstream of a bottleneck, which produced the kind of discontinuity that Edie had earlier identified. The Drake et al. approach was to try to fit the data as they had been obtained, without considering whether there was a portion of the data that was missing. They had intentionally tried to obtain data from as wide a range as possible, but as discussed above it may not be possible to get data from all three parts of the curve at one location.

The third issue is that identified by Cassidy, relating to the time period for collection of the data. The Drake et al. data came from standard loop detectors, working on fixed time intervals. As a consequence, there will be some measurement error, which may well affect the estimation of the bivariate relationships.

The fourth issue is the relationship between car-following models (see Chapter 5) and the models tested by Drake et al. They noted that four of the models they tested "have been shown to be directly related to specific car-following rules", and cited articles by Gazis and co-authors (1959; 1961). The interesting question to raise in the context of the overall appraisal of the Drake et al. results is whether those results raise doubts about the validity of the car-following models for freeways. The car-following models gave rise to four of the speed-density models tested by Drake et al. The results of the testing suggest that the speed-density models are not

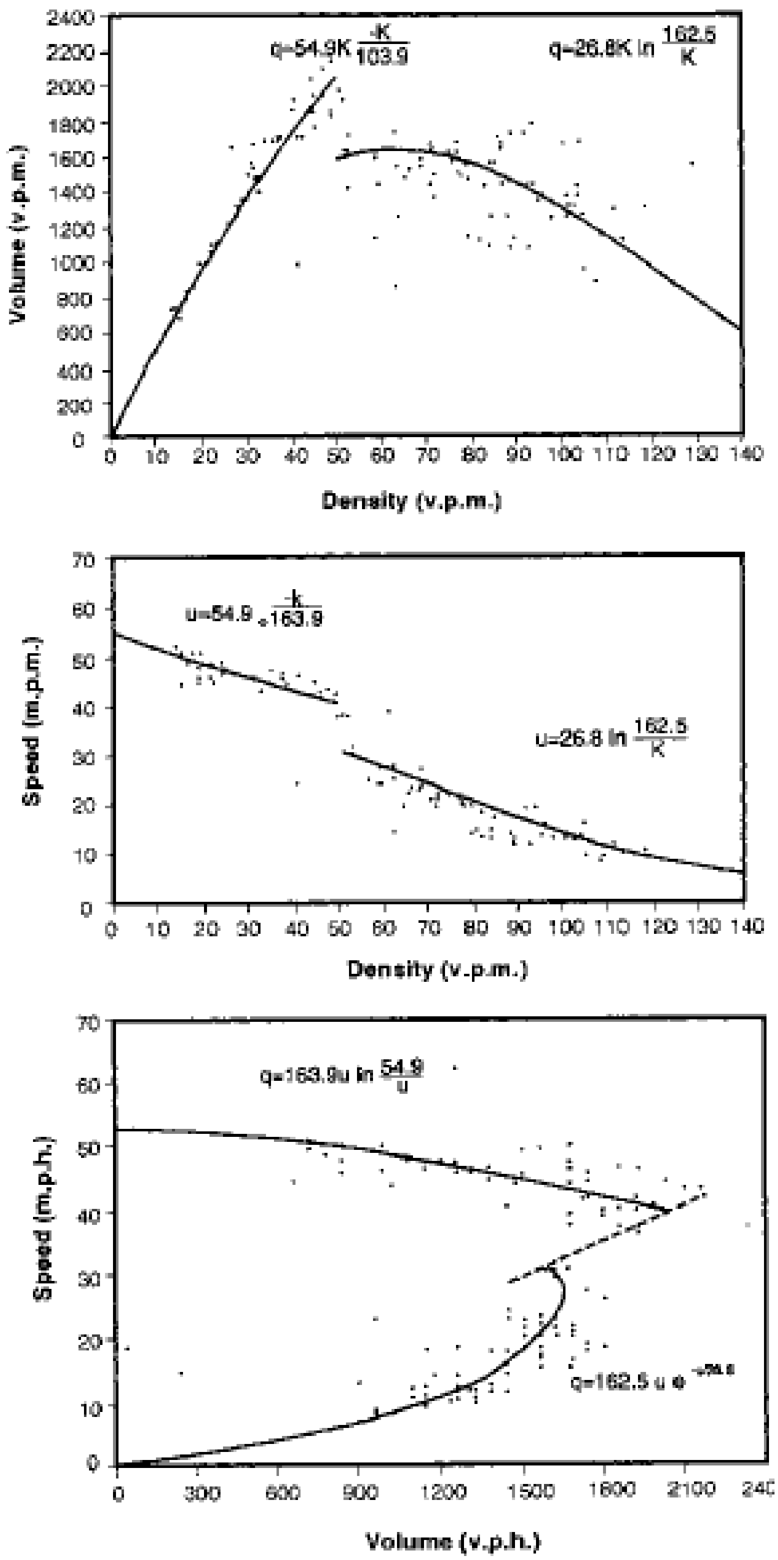

Figure 2.13
Three Parts of Edie's Hypothesis for the Speed-Density Function, Fitted to Chicago Data (Drake et al. 1967).

particularly good. *Modus tollens* in logic says that if the consequences of a set of premises are shown to be false, then one (at least) of the premises is not valid. It is possible, then, that the car-following models are not valid for freeways. (This is not surprising, as they were not developed for this context. Nor, it seems, are they used as the basis for contemporary microscopic freeway simulation models.) On the other hand, any of the three issues just identified may be the source of the failure of the models, rather than their development from car-following models.

2.3.3 Flow-concentration models

Although Gerlough and Huber did not give the topic of flow-concentration models such extensive treatment as they gave the speed-concentration models, they nonetheless thought this topic to be very important, as evidenced by their introductory paragraph for the section dealing with these models (p. 55):

> Early studies of highway capacity followed two principal approaches. Some investigators examined speed-flow relationships at low concentrations; others discussed headway phenomena at high concentrations. Lighthill and Whitham [1955] have proposed use of the flow-concentration curve as a means of unifying these two approaches. Because of this unifying feature, and because of the great usefulness of the flow-concentration curve in traffic control situations (such as metering a freeway), Haight [1960; 1963] has termed the flow-concentration curve "the basic diagram of traffic".

Nevertheless, most flow-concentration models have been derived from assumptions about the shape of the speed-concentration curve. This section deals primarily with work that has focused on the flow-concentration relationship directly. Under that heading is included work that uses either density or occupancy as the measure of concentration.

Edie was perhaps the first to point out that empirical flow-concentration data frequently have discontinuities in the vicinity of what would be maximum flow, and to suggest that therefore discontinuous curves might be needed for this relationship. (An example of his type of curve appears in Figure 2.13 above.) This suggestion led to a series of investigations by May and his students (Ceder 1975; 1976; Ceder and May 1976; Easa and May 1980) to specify more tightly the nature and parameters of these "two-regime" models (and to link those parameters to the parameters of car-following models). The difficulty with their resulting models is that the models often do not fit the data well at capacity (with results similar to those shown in Figure 2.14 for Greenshields' single-regime model). In addition, there seems little consistency in parameters from one location to another. Even more troubling, when multiple days from the same site were calibrated, the different days required quite different parameters.

Koshi, Iwasaki and Ohkura (1983) gave an empirically-based discussion of the flow-density relationship, in which they suggested that a reverse lambda shape was the best description of the data (p. 406): "the two regions of flow form not a single downward concave curve... but a shape like a mirror image of the Greek letter lamda [sic] (λ)". These authors also investigated the implications of this phenomenon for car-following models, as well as for wave propagation. Data with a similar shape to theirs appears in Figure 2.13; Edie's equations fit those data with a shape similar to the lambda shape Koshi et al. suggested.

Although most of the flow-concentration work that relies on occupancy rather than density dates from the past decade, Athol suggested its use nearly 30 years earlier (in 1965). His work presages a number of the points that have come out subsequently and are discussed in more detail below: the use of volume and occupancy together to identify the onset of congestion; the transitions between uncongested and congested operations at volumes lower than capacity; and the use of time-traced plots (i.e. those in which lines connected the data points that occurred consecutively over time) to better understand the operations.

After Athol's early efforts, there seems to have been a dearth of efforts to utilize the occupancy data that was available, until the mid-1980s. One paper from that time (Hall et al. 1986) that utilized occupancy drew on the same approach Athol had used, namely the presentation of time-traced plots. Figure 2.15 shows results for four different days from the same location, 4 km upstream of a primary bottleneck. The data are for the left-most lane only (the high-speed, or passing lane), and are for 5-minute intervals. The first point in the time-connected traces is the one that occurred in the 5-minute period after the data-recording system was turned on in the morning. In part D of the Figure, it is clear that operations had already broken down prior to data being recorded. Part C is perhaps the most intriguing: operations move into higher occupancies (congestion) at flows clearly below maximum flows. Although Parts A and B may be taken to confirm the implicit assumption many traffic engineers have that operations pass through capacity prior to breakdown, Part C gives a clear indication that this does not always happen. Even more important, all four parts of Figure 2.15 show that operations do not go through capacity in returning from congested to uncongested conditions. Operations can 'jump' from one branch of the curve to the other, without staying on the curve. (This same result, not surprisingly, occurs for speed-flow data.)

Each of the four parts of Figure 2.15 show at least one data point between the two 'branches' of the usual curve during the return to uncongested conditions. Because these were 5-minute data, the authors recognized that these points might be the result of averaging of data from the two separate branches. Subsequently, however, additional work utilizing 30-second intervals confirmed the presence of these same types of data (Persaud and Hall 1989). Hence there appears to be strong evidence that traffic operations on a freeway can move from one branch of the curve to the other without going all the way around the capacity point. This is an aspect of traffic behaviour that none of the mathematical models discussed above either explain or lead one to expect. Nonetheless, the phenomenon has been at least implicitly recognized since Lighthill and Whitham's (1955) discussion of shock waves in traffic, which assumes instantaneous jumps from one branch to the other on a speed-flow or flow-occupancy curve. As well, queueing models (e.g. Newell 1982) imply that immediately upsteam from the back end of a queue there must be points where the speed is changing rapidly from the uncongested branch of the speed-flow curve to that of the congested branch. It would be beneficial if flow-

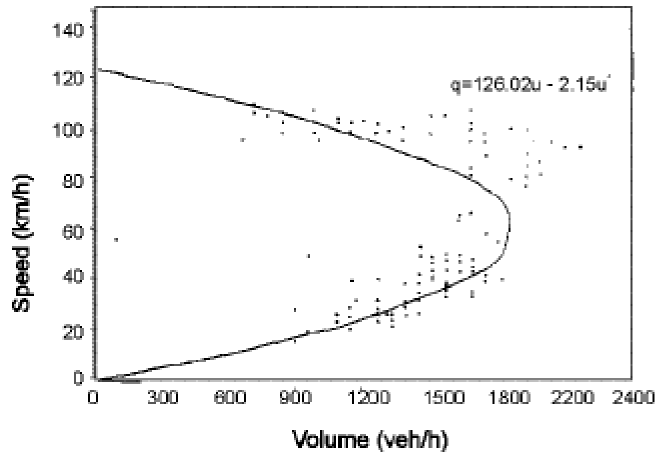

Figure 2.14
Greenshields' Speed-Flow Function Fitted to Chicago Data (Drake et al. 1967).

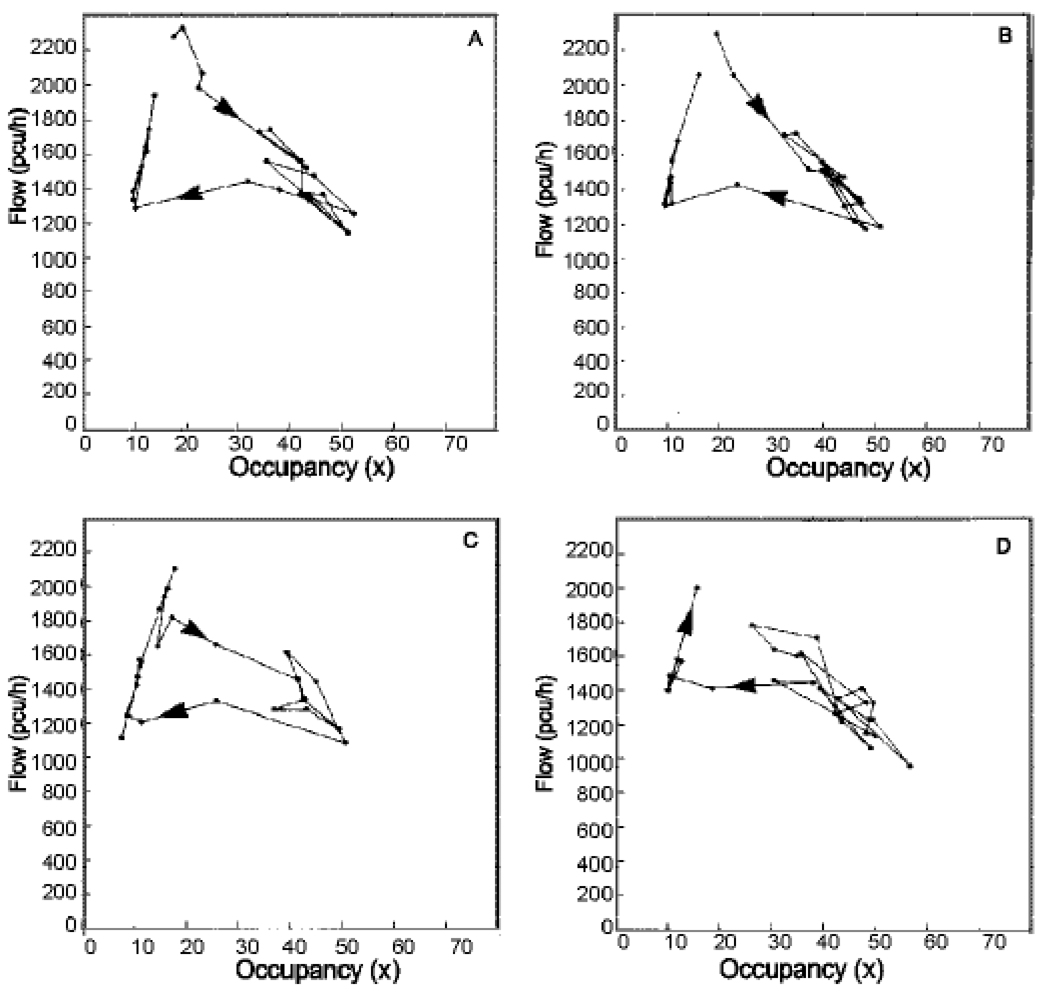

Figure 2.15
Four Days of Flow-Occupancy Data from Near Toronto (Hall et al. 1986).

concentration (and speed-flow) models explicitly took this possibility into account.

One of the conclusions of the paper by Hall et al. (1986) from which Figure 2.15 is drawn is that an inverted 'V' shape is a plausible representation of the flow-occupancy relationship. Although that conclusion was based on limited data from near Toronto, Hall and Gunter (1986) supported it with data from a larger number of stations. Banks (1989) tested their proposition using data from the San Diego area, and confirmed the suggestion of the inverted 'V'. He also offered a mathematical statement of this proposition and a behavioural interpretation of it (p. 58):

> The inverted-V model implies that drivers maintain a roughly constant average time gap between their front bumper and the back bumper of the vehicle in front of them, provided their speed is less than some critical value. Once their speed reaches this critical value (which is as fast as they want to go), they cease to be sensitive to vehicle spacing....

2.4 Three-dimensional models

There has not been a lot of work that attempts to treat all three traffic flow variables simultaneously. Gerlough and Huber presented one figure (reproduced as Figure 2.16) that represented all three variables, but said little about this, other than (1) "The model must be on the three-dimensional surface $u = q/k$," and (2) "It is usually more convenient to show the model of [Figure 2.16] as one or more of the three separate relationships in two dimensions..." (p. 49). As was noted earlier, however, empirical observations rarely accord exactly with the relationship $q=u\ k$, especially when the observations are taken during congested conditions. Hence focusing on the two-dimensional relationships will not often provide even implicitly a valid three-dimensional relationship.

In this context, a paper by Gilchrist and Hall (1989) is interesting because it presents three-dimensional representations of empirical observations, without attempting to fit them to a theoretical representations. The study is limited, in that data from only one location, upstream of a bottleneck, was presented. To enable better visualization of the data, a time-connected trace was used. The projections onto the standard two-dimensional surfaces of the data look much as one might expect. The surprises came in looking at oblique views of the three-dimensional representation, as in Figures 2.17 and 2.18. From one perspective (Figure 2.17), the traditional sideways U-shape that we have been led to expect is quite apparent, and projections of that are easily visualized onto, for example, the speed-flow surface (the face labeled with a 3, on the left side of the ‘box’). From a different perspective (Figure 2.18) that shape is hardly apparent at all, although indications of an inverted 'V' can be seen, which would project onto a flow-occupancy plot on face 1 of the box.. Black and white were alternated for five speed ranges. In both figures, the dark lines at the left represent the data with speeds above 80 km/hr. The light lines closest to the left cover the range 71 to 80 km/hr; the next, very small, area of dark lines contains the range 61 to 70 km/hr; the remaining light lines represent the range 51 to 60 km/hr; and the dark lines to the right of the diagram represent speeds below 50 km/hr.

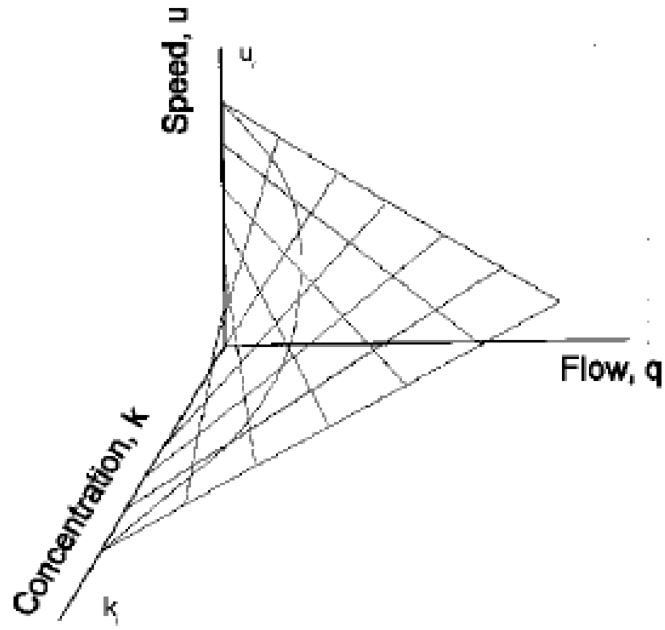

Figure 2.16
The Three-Dimensional Surface for Traffic Operations, as in Transportation Research Board Special Report 165 (Gerlough and Huber 1975).

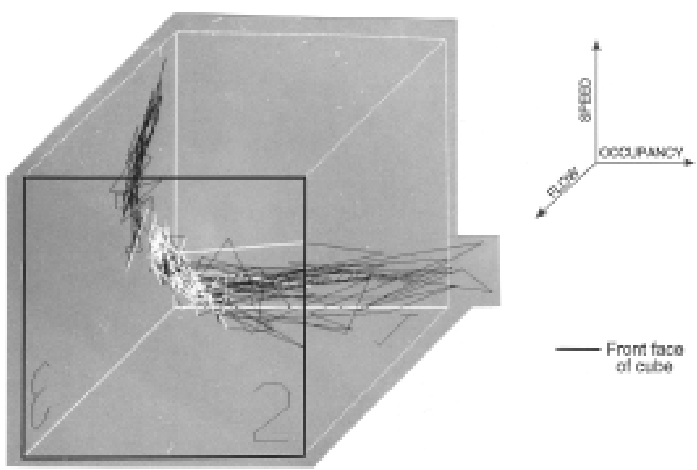

Figure 2.1 7
One Perspective on the Three-Dimensional Speed-Flow-Concentration Relationship (Gilchrist and Hall 1989).

One recent approach to modelling the three traffic operations variables directly has been based on the mathematics of catastrophe theory. (The name comes from the fact that while most of the variables being modelled change in a continuous fashion, at least one of the variables can make sudden discontinuous changes, referred to as catastrophes by Thom (1975), who originally developed the mathematics for seven such models, ranging from two dimensions to eight.) The first effort to apply these models to traffic data was that by Dendrinos (1978), in which he suggested that the two-dimensional catastrophe model could represent the speed-flow curve. A more fruitful model was proposed by Navin (1986), who suggested that the three-dimensional 'cusp' catastrophe model was appropriate for the three traffic variables. The feature of the cusp catastrophe surface that makes it of interest in the traffic flow context is that while two of the variables (the control variables) exhibit smooth continuous change, the third one (the state variable) can undergo a sudden 'catastrophic' jump in its value. Navin suggested that speed was the variable that underwent this catastrophic change, while flow and occupancy were the control variables.

While Navin's presentation was primarily an intuitive one, without recourse to data, Hall and co-authors picked up on the idea and attempted to flesh it out both mathematically and empirically. The initial effort appears in Hall (1987), in which the basic idea was presented, and some data applied. Figure 2.19 is a representation, showing the partially-folded (and torn) surface that is the Maxwell version of the cusp catastrophe surface, approximately located traffic data on that surface, and axes external to the surface representing the general correspondence with traffic variables. Further elaboration of this model is provided by Persaud and Hall (1989). Acha-Daza and Hall (1993) compared the ability of this model with the ability of some of the earlier models discussed above, to estimate speeds from flows obtained using inductive loop detectors and 30-second data, and found the catastrophe theory model to be slightly better than they were. Despite its empirical success, the problem with the model is that it appears to be inconsistent with the basic definitions and relationships with which this chapter opened.

2.5 Summary and links to other chapters

The current status of mathematical models for speed-flow-concentration relationships is in a state of flux. The models that dominated the discourse for nearly 30 years are incompatible with the data currently being obtained, and with currently accepted depictions of speed-flow curves, but no replacement models have yet been developed. The analyses of Cassidy and Newell have shown that the data used to develop most of the earlier models were flawed, as has been described above. An additional difficulty was noted by Duncan (1976; 1979): transforming variables, fitting equations, and then transforming the equations back to the original variables can lead to biased results, and is very sensitive to small changes in the initial curve-fitting. Progress has been made in understanding the relationships among the key traffic variables, but there is considerable scope for better models still.

It is important to note that the variables and analyses that have been discussed in this chapter are closely related to topics in several of the succeeding chapters. The human factors discussed in Chapter 3, for example, help to explain some of the variability in the data that have

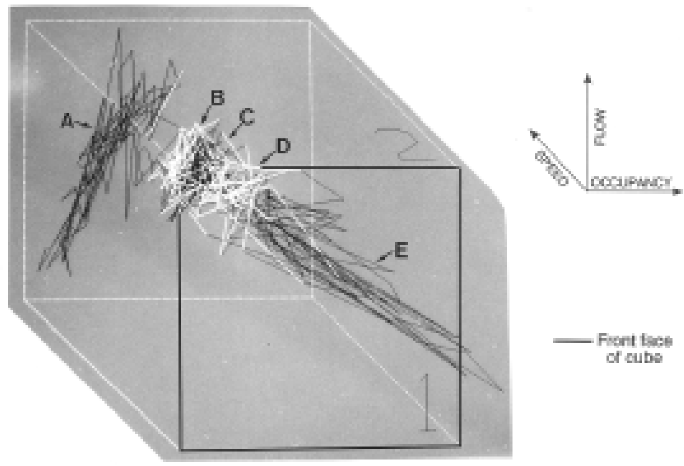

Figure 2.1 8
Second Perspective on the Three-Dimensional Relationship (Gilchrist and Hall 1989).

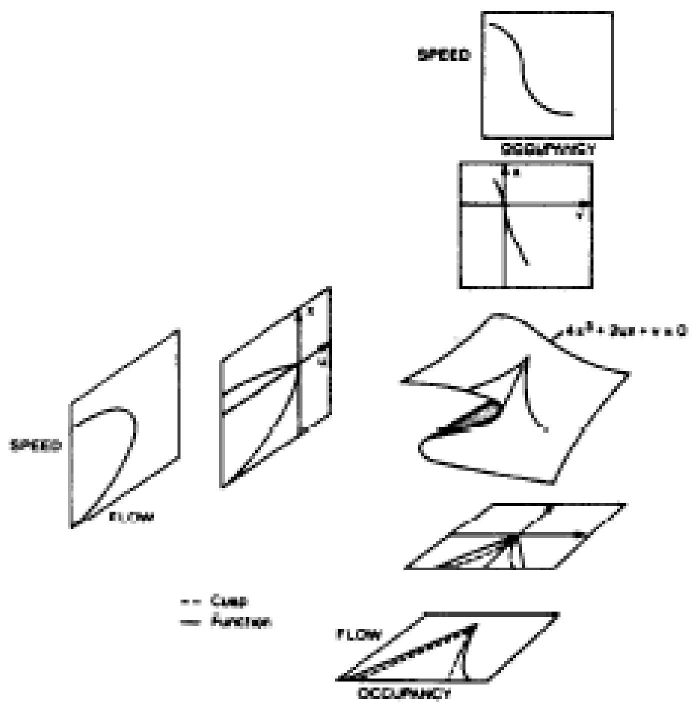

Figure 2.19
Catastrophe Theory Surface Showing Sketch of a Possible Freeway Function, and Projections and Transformations from That (Hall 1987).

been discussed here. As mentioned earlier, some of the bivariate models discussed above have been derived from car-following models, which are covered in Chapter 4. In addition, brief mention was made here of Lighthill and Witham's work, which has spawned a large literature related to the behavior of shocks and waves in traffic flow, covered in Chapter 5. All of these topics are inter-related, but have been addressed separately for ease of understanding.

Acknowledgements

The assistance of Michael Cassidy in developing this latest revision of this chapter was very helpful, as noted in the preface to the chapter. Valuable comments on an early draft of this material were received from Jim Banks, Frank Montgomery and Van Hurdle. The assistance of Richard Cunard and the TRB in bringing this project to completion is also appreciated.

References

Acha-Daza, J.A. and Hall, F.L. 1993. A graphical comparison of the predictions for speed given by catastrophe theory and some classic models, *Transportation Research Record* **1398**, 119-124.

Agyemang-Duah, K. and Hall, F.L. 1991. Some issues regarding the numerical value of freeway capacity, in *Highway Capacity and Level of Service*, Proceedings of the International Symposium on Highway Capacity, Karlsruhe. U. (Brannolte, ed.) Rotterdam, Balkema, 1-15.

Athol, P. 1965. Interdependence of certain operational characteristics within a moving traffic stream. *Highway Research Record* **72**, 58-87.

Banks, J.H. 1989. Freeway speed-flow-concentration relationships: more evidence and interpretations, *Transportation Research Record* **1225**, 53-60.

Banks, J.H. 1990. Flow processes at a freeway bottleneck. *Transportation Research Record* **1287**, 20-28.

Banks, J.H. 1991a. Two-capacity phenomenon at freeway bottlenecks: A basis for ramp metering?, *Transportation Research Record* **1320**, 83-90.

Banks, J.H. 1991b. The two-capacity phenomenon: some theoretical issues, *Transportation Research Record* **1320**, pp. 234-241.

Cassidy, MJ 1998 Bivariate relations in nearly stationary highway traffic. *Transp Res B, 32B,* 49-59.

Cassidy, MJ 1999 Traffic flow and capacity, in *Handbook of transportation science*, RW Hall (ed). Norwell, MA: Kluwer Academic Publishers.

Cassidy , MJ and Coifman, B 1997 Relation among average speed, flow and density and the analogous relation between density and occupancy. *Transportation Research Record 1591*, 1-6.

Ceder, A. 1975. Investigation of two-regime traffic flow models at the micro- and macro-scopic levels. Ph.D. dissertation, University of California, Berkeley, California.

Ceder, A. 1976. A deterministic traffic flow model for the two-regime approach. *Transportation Research Record* **567**, 16-32.

Ceder, A. and May, A.D. 1976. Further evaluation of single- and two-regime traffic flow models. *Transportation Research Record* **567**, 1-15.

Chin, H.C. and May, A.D. 1991. Examination of the speed-flow relationship at the Caldecott Tunnel, *Transportation Research Record* **1320**.

Daganzo, CF 1997 *Fundamentals of transportation and traffic operations*. New York: Elsevier.

Dendrinos, D.S. 1978. Operating speeds and volume to capacity ratios: the observed relationship and the fold catastrophe, *Transportation Research Record* **12**, 191-194.

Drake, J.S., Schofer, J.L., and May, A.D. 1967. A Statistical Analysis of Speed Density Hypotheses, *Highway Research Record* **154**, 53-87.

Duncan, N.C. 1976. A note on speed/flow/concentration relations, *Traffic Engineering and Control*, 34-35.

Duncan, N.C. 1979. A further look at speed/flow/concentration, *Traffic Engineering and Control*, 482-483.

Easa, S.M. and May, A.D. 1980. Generalized procedures for estimating single- and two-regime traffic flow models, *Transportation Research Record* **772**, 24-37.

Edie, L.C. 1961. Car following and steady-state theory for non-congested traffic, *Operations Research* **9**, 66-76.

Edie, LC 1965 Discussion of traffic stream measurements and definitions. *Proc Int Symp. On the Theory of Traffic Flow* J. Almond (ed) Paris: OECD, 139-154.

Edie, LC 1974 Flow theories, in *Traffic Science* DC Gazis (ed) New York: Wiley, 8-20.

Edie, LC and Foote, R 1960 Effect of shock waves on tunnel traffic flow. *Proc Highway Res Board 39*, 492-505.

Edie, L.C., Herman, R., and Lam, T., 1980. Observed multilane speed distributions and the kinetic theory of vehicular traffic, *Transportation Science* **14**, 55-76.

Gazis, D.C., Herman, R., and Potts, R. 1959. Car-following theory of steady-state traffic flow. *Operations Research* **7**, 499-505.

Gazis, D.C., Herman, R., and Rothery, R.W. 1961. Nonlinear follow-the-leader models of traffic flow. *Operations Research* **9**, 545-567.

Gerlough, D.L. and Huber, M.J. 1975. Traffic Flow Theory: a monograph. Special Report 165, Transportation Research Board (Washington D.C.: National Research Council)

Gilchrist, R. 1988. *Three-dimensional relationships in traffic flow theory variables*. Unpublished Master's degree report, Dept. of Civil Engineering, McMaster University, Hamilton, Ontario, Canada.

Gilchrist, R.S. and Hall, F.L. 1989. Three-dimensional relationships among traffic flow theory variables, *Transportation Research Record* **1225**, 99-108.

Greenberg, H. 1959. An analysis of traffic flow, *Operations Research*, **Vol. 7**, 78-85.

Greenshields, B.D. 1935. A study of traffic capacity, *HRB Proceedings* **14**, 448-477.

Haight, F.A. 1960. The volume, density relation in the theory of road traffic, *Operations Research* **8**, 572-573.

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

Hall, F.L. 1987. An interpretation of speed-flow-concentration relationships using catastrophe theory. *Transportation Research A*, **21A**, 191-201.

Hall, F.L., Allen, B.L. and Gunter, M.A. 1986. Empirical analysis of freeway flow-density relationships, *Transportation Research A*, **20A**, 197-210.

Hall, F.L. and Gunter, M.A. 1986. Further analysis of the flow-concentration relationship, *Transportation Research Record* **1091**, 1-9.

Hall, F.L. and Hall, L.M. 1990. Capacity and speed flow analysis of the QEW in Ontario" *Transportation Research Record* **1287**, 108-118.

Hall, F.L, Hurdle, V.F. and Banks, J.H. 1992. Synthesis of recent work on the nature of speed-flow and flow-occupancy (or density) relationships on freeways. In *Transportation Research Record* **1365**, TRB, National Research Council, Washington, D.C. 12-18.

Heidemann, D. and R. Hotop, R. 1990. Verteilung der Pkw-Geschwindigkeiten im Netz der

Bundesautobahnen- Modellmodifikation und Aktualisierung. *Straße und Autobahn* - Heft 2, 106-113.

HCM 1965. *Special Report 87: Highway Capacity Manual*. Transportation Research Board.

HCM 1985. *Special Report 209: Highway Capacity Manual*. Transportation Research Board.

Hsu, P. and Banks, J.H. 1993. Effects of location on congested-regime flow-concentration relationships for freeways, *Transportation Research Record* **1398**, 17-23.

Huber, M.J. 1957. Effect of temporary bridge on parkway performance. *Highway Research Board Bulletin* **167** 63-74.

Hurdle, V.F. and Datta, P.K. 1983. Speeds and flows on an urban freeway: some measurements and a hypothesis. *Transportation Research Record* **905**, 127-137.

Koshi, M., Iwasaki, M. and Okhura, I. 1983. Some findings and an overview on vehicular flow characteristics. *Proceedings*, 8*th International Symposium on Transportation and Traffic Flow Theory* (Edited by Hurdle, V.F., Hauer, E. and Steuart, G.F.) University of Toronto Press, Toronto, Canada, 403-426.

Kühne, R. and Immes, S. 1993. Freeway control systems for using section-related traffic variable detection. *Proceedings of the Pacific Rim TransTech Conference*, Seattle. New York: Am. Soc. Civil Eng., 56-62.

Lighthill, M.J., and Whitham, G.B. 1955. On kinematic waves: II. A theory of traffic flow on long crowded roads. *Proceedings of the Royal Society* A229 No. 1178, 317-145.

Makagami, Y., Newell, G.F. and Rothery, R. 1971. Three-dimensional representation of traffic flow, *Transportation Science*, Vol. 5, No. 3, pp. 302-313.

May, A.D. 1990. *Traffic Flow Fundamentals*. Prentice-Hall.

May, A.D., Athol, P., Parker, W. and Rudden, J.B. 1963. Development and evaluation of Congress Street Expressway pilot detection system, *Highway Research Record* **21**, 48-70.

Morton, T.W. and Jackson, C.P. 1992. Speed/flow geometry relationships for rural dual-carriageways and motorways. TRRL Contractor Report 279. 55 pages. Transport and Road Research Laboratory, Crowthorne, Berkshire.

Moskowitz, K 1954 Waiting for a gap in a traffic stream *Proc Highway Res Board 33*, 385-395.

Navin, F. 1986. Traffic congestion catastrophes. *Transportation Planning Technology* **11**, 19-25.

Newell G.F. 1971 *Applications of Queueing Theory*. London: Chapman Hall.

Newell, G.F. 1982. *Applications of Queueing Theory*, second edition. London: Chapman and Hall.

Newell, G.F. 1993. A simplified theory of kinematic waves in highway traffic, *Transportation Research B*, 27*B*: Part I, General theory, 281-287; Part II, Queueing at freeway bottlenecks, 289-303; Part III, Multi-destination flows, 305-313.

Newell, GF unpublished Notes on transportation operations Berkeley, CA: Univesity of California.

Persaud, B.N. and Hall, F.L. 1989. Catastrophe theory and patterns in 30-second freeway data -- implications for incident detection, *Transportation Research A* **23A**, 103-113.

Persaud, B.N. and Hurdle, V.F. 1988. Some new data that challenge some old ideas about speed-flow relationships. *Transportation Research Record* **1194**, 191-198.

Ringert, J. and Urbanik, II, T. 1993. Study of freeway bottlenecks in Texas, *Transportation Research Record* **1398**, 31-41.

Robertson, D. 1994. *Manual of Transportation Engineering Studies*, fifth edition. Arlington, VA: Institute of Transportation Engineers.

Stappert, K.H. and Theis, T.J. 1990. Aktualisierung der Verkehrsstärke- /Verkehrsgeschwindigkeitsbeziehungen des BVWP-Netzmodells. Research Report VU18009V89 Heusch/Boesefeldt, Aachen.

Thom, R. 1975. *Structural Stability and Morphogenesis*. English translation by D.H. Fowler of *Stabilite Structurelle et Morphogenese*. Benjamin: Reading, Massachusetts.

Underwood, R.T. 1961. Speed, volume, and density relationships: Quality and theory of traffic flow. Yale Bureau of Highway Traffic, pp. 141-188, as cited in Drake et al. 1967.

Wardrop, J.G. and Charlesworth, G. 1954. A method of estimating speed and flow of traffic from a moving vehicle. Proceedings of the Institution of Civil Engineers, Part II, Vol. 3, 158-171.

Wattleworth, J.A. 1963. Some aspects of macroscopic freeway traffic flow theory, *Traffic Engineering* **34**: 2, 15-20.

Wemple, E.A., Morris, A.M. and May, A.D. 1991. Freeway capacity and level of service concepts, in *Highway Capacity and Level of Service*, Proc. of the International Symposium on Highway Capacity, Karlsruhe (U. Brannolte, Ed.) Rotterdam: Balkema, 439-455.

Wright, C. 1973. A theoretical analysis of the moving observer method, *Transportation Research* **7**, 293-311.