6 TRAFFIC FLOW AND CAPACITY

Michael J. Cassidy

6.1 Introduction

The design of highways, runways, ports or any transportation facility is guided by knowledge and theory of the traffic streams they serve. A facility's scale, its geometry and its control measures are selected to affect certain properties of its traffic, such as the travel delay, the separation between vehicles, etc. In the case of highway traffic, the emphasis of this chapter, these are usually properties that are collected from, or averaged over, some number of vehicles. This is because the behavior of one driver differs from that of another, sometimes in complicated or even unexpected ways, and the traffic engineer typically seeks properties that are reproducible or predictable; i.e., properties that are not sensitive to driver variations.

Chapter 6 is devoted to methods of measuring traffic stream properties and of predicting how these properties evolve over time and space. Certain emphasis is given to flow restrictions, or bottlenecks, and to the estimation of their capacities since traffic streams are often impacted by these restrictions.

Section 6.1 provides some important definitions along with descriptions of some graphical tools for analyzing the motion of objects on transport systems. Most of what is presented here is applicable to any mode of transportation. Moreover, this information is necessary background for the treatment of highway traffic offered in the remaining sections of the chapter. Section 6.2 describes methods of processing traffic data measured, for example, by loop detectors to identify bottleneck locations along highway facilities without traffic signals or other exogenous controls. The use of these methods to estimate bottleneck capacities is likewise shown here. Section 6.3 presents methods of estimating capacities and vehicle delays at highway intersections controlled by traffic signals or stop signs. Theories for predicting the evolution of highway traffic are the subject of section 6.4. A simple theory is described here in some detail and other theories are briefly noted.

The chapter provides references for all of the topics covered. Notes on the historical developments and future research directions are likewise included for many of the subjects.

Basic Concepts

This first section includes definitions for some of the properties commonly used to characterize traffic streams. So-called generalized definitions, which preserve useful relations between the properties, are part of this discussion. Also described in section 6.1 is a three-dimensional representation of traffic streams. This representation makes clear the conservation concepts that are fundamental to theories of traffic evolution. In particular, it illustrates the relation between two important graphical tools for presenting and interpreting traffic data: 1) curves of cumulative vehicle count and 2) trajectories plotted on time-space diagrams. A description of the latter tool is the starting point for this section.

Before embarking on this discussion, however, there are two points that deserve mention. First, the subjects covered in section 6.1 do not involve theory or conjecture. Rather the concepts are true by definition. Secondly, the discussion in this section owes much to notes composed by Newell (unpublished) for a graduate course in transportation engineering and to a book written by Daganzo (1997).

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 6-1 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, $x_\ell(t)$.

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."

The literature includes numerous illustrations of how these diagrams, even when drawn approximately, can be used in solving problems that frequently arise in transport. As examples, Daganzo (1997) shows how trajectory plots can help to select desirable scheduling policies in rail and in sea transportation; Newell (1979) used them in deriving expressions of airport runway capacity; and they are a widely used tool for synchronizing traffic signals along an arterial (Newell, 1989).

This chapter will frequently rely upon time-space diagrams to illustrate fundamental concepts. They are used immediately below to convey the precise meanings of some important properties of the traffic stream.

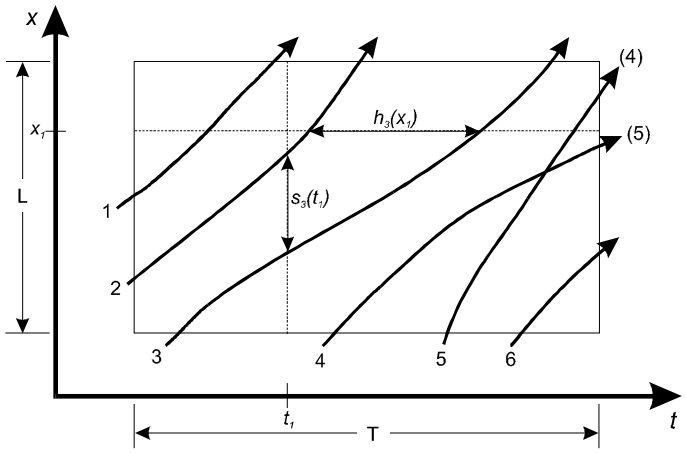

Figure 6-1. Time-space diagram.

Definitions of Some Traffic Stream Properties. It is evident from Figure 6-1 that the slope of the ℓth trajectory is object ℓ's instantaneous velocity, $v_\ell(t)$, i.e.,

$$v_\ell(t) \equiv dx_\ell(t)/dt\ , \tag{6.1}$$

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 6-1, 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.,

$$h_i(x_1) \equiv t_i(x_1) - t_{i-1}(x_1) \tag{6.2}$$

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

$$q(T, x_1) \equiv m/T. \tag{6.3}$$

For observation intervals containing large m,

$$\sum_{i=1}^{m} h_i(x_1) \approx T \tag{6.4}$$

and thus,

$$q(T, x_1) \approx \frac{1}{\frac{1}{m}\sum_{i=1}^{m} h_i(x_1)} = \frac{1}{\bar{h}(x_1)}, \tag{6.5}$$

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.,

$$s_j(t_1) \equiv x_{j-1}(t_1) - x_j(t_1). \tag{6.6}$$

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.,*

$$k(L, t_1) \equiv n/L. \tag{6.7}$$

If the *L* contains large *n*,

$$\sum_{j=1}^{n} s_j(t_1) \approx L \tag{6.8}$$

and

$$k(L, t_1) \approx \frac{1}{\frac{1}{n}\sum_{j=1}^{n} s_j(t_1)} = \frac{1}{\bar{s}(t_1)}, \tag{6.9}$$

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

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*,

$$\alpha(T, x_1) = \frac{1}{m}\sum_{i=1}^{m} \alpha_i(x_1), \tag{6.10}$$

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

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

$$\alpha(L, t_1) = \frac{1}{n} \sum_{j=1}^{n} \alpha_j(t_1). \tag{6.11}$$

If, for example, α is spacing, $\alpha(L, t_1)$ 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 $L \times T$ 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.

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 6-2, 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 6-2 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 6-2, 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.

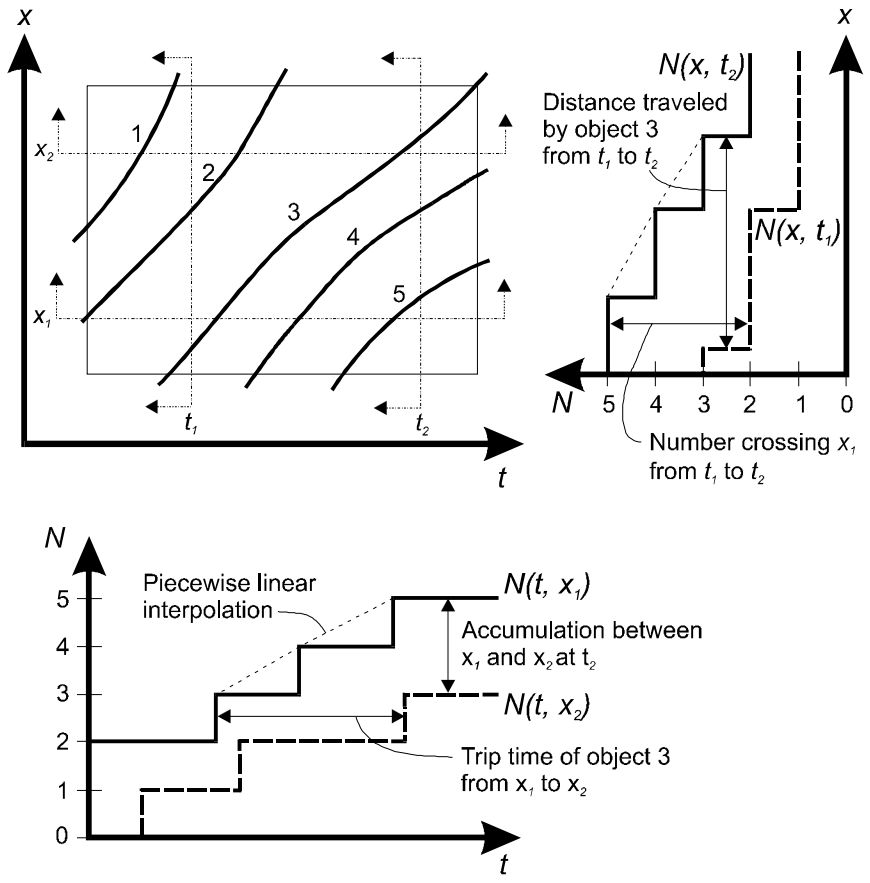

Figure 6-2. Three-dimensional representation.

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). 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.

By examining trends on the cumulative curves, one can observe how flows and densities change with time and space, respectively. This can be a powerful diagnostic and examples are provided in later sections. Suffice to say that by defining flows and densities as they are displayed on the cumulative curves, their values may be taken over intervals that exhibit fixed trends (i.e., 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).

The Conservation Law. The existence of the surface $N(t, x)$ implies that objects did not enter or exit within the region of interest. If the N can be replaced by a smooth surface N' so that at all points within the region the *instantaneous* flows and densities can be defined as $\partial N'(t,x)/\partial t$ and $\partial N'(t,x)/\partial x$, respectively, and if N' has second derivatives (i.e., flow and density are smooth), then $\partial^2 N'(t,x)/\partial x \partial t$ must be equal to $\partial^2 N'(t,x)/\partial t \partial x$ and thus

$$\frac{\partial q(t,x)}{\partial x} = -\frac{\partial k(t,x)}{\partial t}. \tag{6.12a}$$

The more common form of this conservation equation is

$$\frac{\partial q(t,x)}{\partial x} + \frac{\partial k(t,x)}{\partial t} = 0. \tag{6.12b}$$

It is by direct consequence of the conservation equation that the speed of an interface separating two (different) stationary traffic conditions, u, is the change in flow across the interface over the change in density across the interface; i.e.,

$$u = \Delta q / \Delta k \,. \tag{6.13}$$

The reader may refer to Daganzo (1997, pp. 97-103) for the simple derivation of (6.13) and for further discussion of this. The conservation equation also gives rise to the well-known expression for computing the relative flow measured by a moving observer in (stationary) traffic; see again Daganzo (1997).

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 6-3 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 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 6-3, Edie's generalized

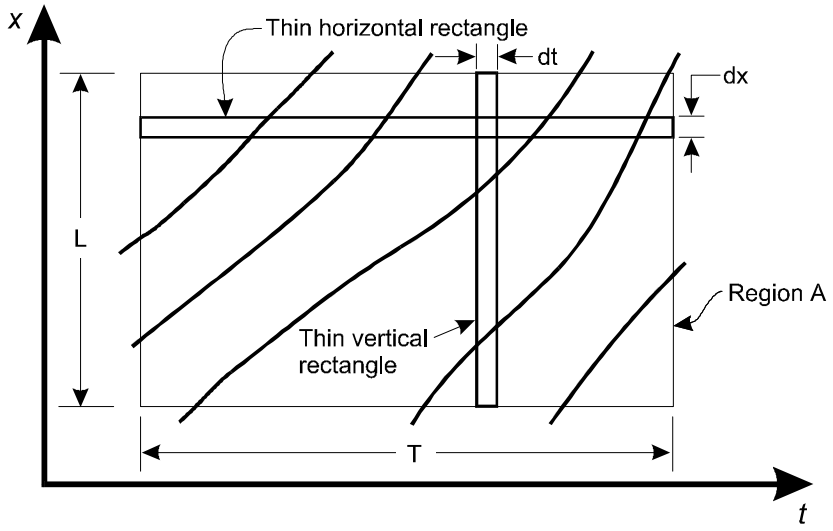

Figure 6-3. Trajectories in time-space region.

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 6-3 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, $d(A) = m \cdot dx$, the generalized mean velocity becomes

$$v(A) = \frac{d(A)}{t(A)} = \frac{1}{\frac{1}{m}\sum_{i=1}^{m}\frac{1}{v_i}}, \tag{6.14}$$

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

$$v(A) = \left[\frac{1}{m}\sum_{i=1}^{m} p_i\right]^{-1}. \tag{6.15}$$

Eq. 6.15 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).

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 6-4. 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 $k(A)$,

$$average\ vehicle\ length = \frac{\rho(A)}{k(A)} = \frac{area\ of\ the\ shaded\ strips}{|A|} \cdot \frac{|A|}{t(A)}. \tag{6.16}$$

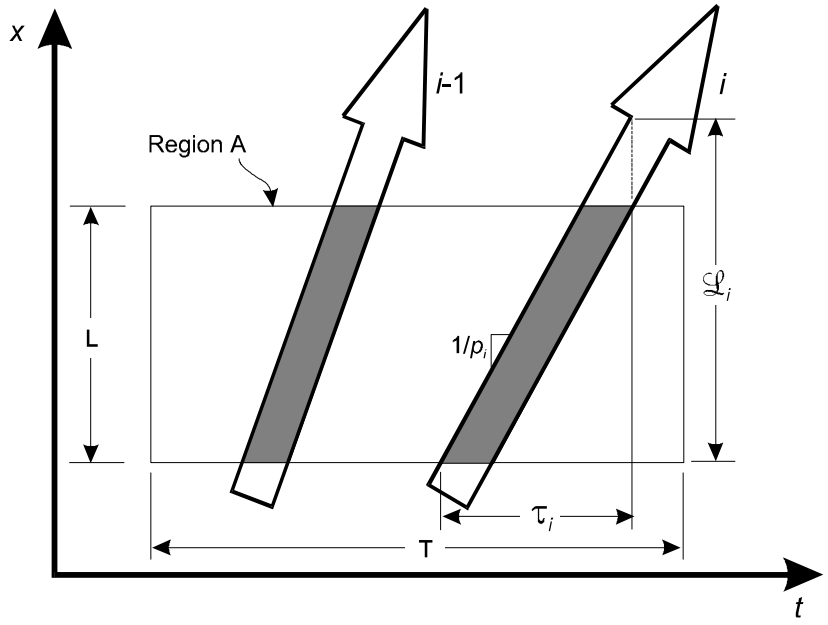

Figure 6-4. Trajectories of vehicle fronts and rears.

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, $\mathscr{L}_i$ is the summed length of the detection zone and the length of vehicle i. Therefore,

$$\sum_{i=1}^{m} \tau_i = \sum_{i=1}^{m} \mathscr{L}_i \cdot \frac{1}{v_i} = \sum_{i=1}^{m} \mathscr{L}_i p_i \tag{6.17}$$

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

$$\frac{\sum_{i=1}^{m} \tau_i}{T} = \frac{\frac{1}{m} \cdot \sum_{i=1}^{m} \tau_i}{\frac{1}{m} \cdot T} = q(A) \cdot \frac{1}{m} \cdot \sum_{i=1}^{m} \tau_i, \tag{6.18}$$

it follows that

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

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

$$\rho = k(A) \cdot \left[\frac{\sum_{i=1}^{m} \mathcal{L}_i \cdot p_i}{\sum_{i=1}^{m} p_i} \right], \tag{6.19}$$

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 (6.19) 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 $\mathcal{L}_i$ 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.

6.2 Freeway Bottlenecks and their Capacities

This section provides description of some simple diagnostics for locating bottlenecks on freeways and on highways without control devices, such as traffic signals. Also described here are techniques for estimating the capacities of these bottlenecks. Cumulative curves constructed from counts and occupancies measured at neighboring locations along the roadway serve as the diagnostics. By transforming and visually inspecting these curves as described below, one can verify the occurrence of an active bottleneck, where the word active is used to denote 1) a queue's presence immediately upstream, which ensures that vehicles are discharging through the bottleneck at a maximum rate, and 2) the absence of any downstream effects that would impede this discharge. Once having identified these two essential conditions, the bottleneck's capacity may be estimated. Plotting the cumulative counts that have been measured immediately downstream of the bottleneck can aid in this endeavor.

To illustrate these diagnostics, they are applied to a bottleneck that formed downstream of a merge; some details of this site are described below. Of note, the diagnostics may be used for examining bottlenecks caused by other types of

geometric inhomogeneities, including curves, lane reductions, and diverges.[1] They can also be applied to bottlenecks formed by incidents, such as vehicle stalls or collisions.

An Example Application

The diagnostics will be described with data taken from the freeway section shown in Figure 6-5, a segment of the Queen Elizabeth Way in Ontario, Canada. All the data presented here came from a single morning rush and were measured with loop detectors installed at four locations along this freeway segment. These four detectors are labeled in the figure as per the numbering strategy that had been adopted by the region's transportation authority. The vehicle counts and occupancies were measured in 30-second intervals and the resulting step-wise cumulative curves were smoothed using piece-wise linear interpolations.

As a useful aside, cumulative count curves may be obtained from vehicle arrival times measured by human observers stationed roadside when loop detectors are not deployed near a bottleneck of interest. A detailed description of one such experiment can be found in Smilowitz, et al. (1998).

Locating the Active Bottleneck

Figure 6-6 shows cumulative count, or N-curves. These were constructed from counts measured across all lanes at the four detector stations during the onset of queueing. The counts at each detector were started ($N = 0$) with the passage of an imaginary reference vehicle. These passage times were based upon estimated free-flow trip times between each detector because vehicle $N \approx 0$ did not encounter queueing at the site. The curves in Figure 6-6 have been transformed in the following two ways.

First, each curve, along with its respective time axis, was shifted to the right by the average free flow trip time between the respective detector and downstream detector 25. Having done this, the vertical displacements between curves are the excess vehicular accumulations due to traffic delays. Such a shift is advantageous because two superimposed curves indicate that traffic in the intervening segment is flowing freely; every feature of an upstream N-curve is passed to its downstream neighbor a free-flow trip time later. Secondly, Figure 6-6 shows only the differences between each curve of cumulative count to time t and the line $N = q_o t'$,

[1] At a diverge bottleneck, the queue(s) formed by vehicles wishing to exit the freeway may entrap some through-moving vehicles and thereby affect the bottleneck's capacity. The reader may note for now these special circumstances that surround diverge operation. More is said about this in section 6.4.

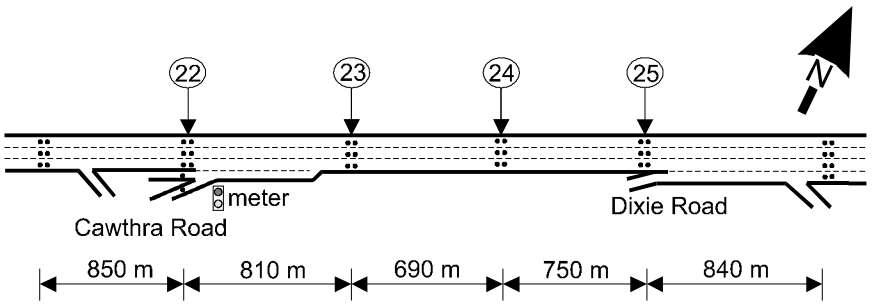

Figure 6-5. Segment of Queen Elizabeth Way.

where t' is the elapsed time from the curve's starting point (N=0) and q_o is the rate used for re-scaling the cumulative curve. This is important because reducing the N (displayed by the curves) by a background flow q_o magnifies details without changing the excess accumulations (Cassidy and Windover, 1995).

The superimposed curve portions in this figure indicate that traffic was initially in free flow and remained in free flow between detectors 24 and 25. The marked separation of curves 24/25 from curve 23 from about 6:27 onward (as shown by the darkened arrow) indicates that a bottleneck was activated a little earlier between detectors 24 and 23. The subsequent separation of curve 23 from curve 22 indicates when the queue arrived to detector 23.

This illustrates how transformed cumulative count curves expose active bottlenecks by revealing the excess accumulations upstream and the free-flow conditions downstream. Further verification of a bottleneck's activation can be obtained by using re-scaled curves of cumulative occupancy, as described below.

Additional evidence of the bottleneck. A bottleneck's location may be confirmed using curves of cumulative occupancy versus time (T-curves) where cumulative occupancy is the total vehicular trip time over the loop detector by time t (Lin and Daganzo, 1997). To illustrate this, Figure 6-7 presents T-curves for the four detector stations. As before, these curves were constructed for times near the onset of queueing and the occupancies were those measured in all travel lanes. Again for the purpose of magnifying details, the figure presents the differences between each curve of cumulative occupancy to time t and the line $T = b_o(x)t'$, where $b_o(x)$ is the

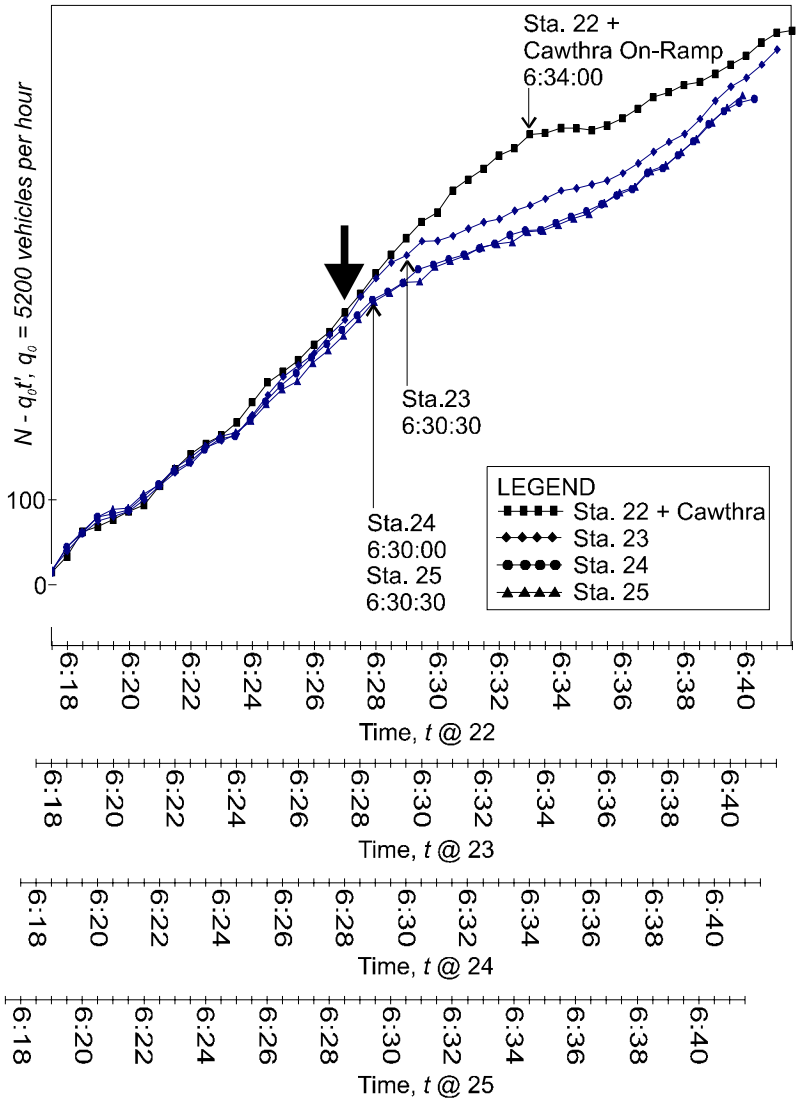

Figure 6-6. Transformed N-curves.

background occupancy rate used at detector x and t' is the elapsed time from the curve's starting point ($T = 0$).

The T-curves in the lower half of Figure 6-7 display concave shapes, indicating sudden reductions in the occupancy rates at detectors 24 and 25. These lower occupancies prevailed during times that coincided (approximately) with the flow

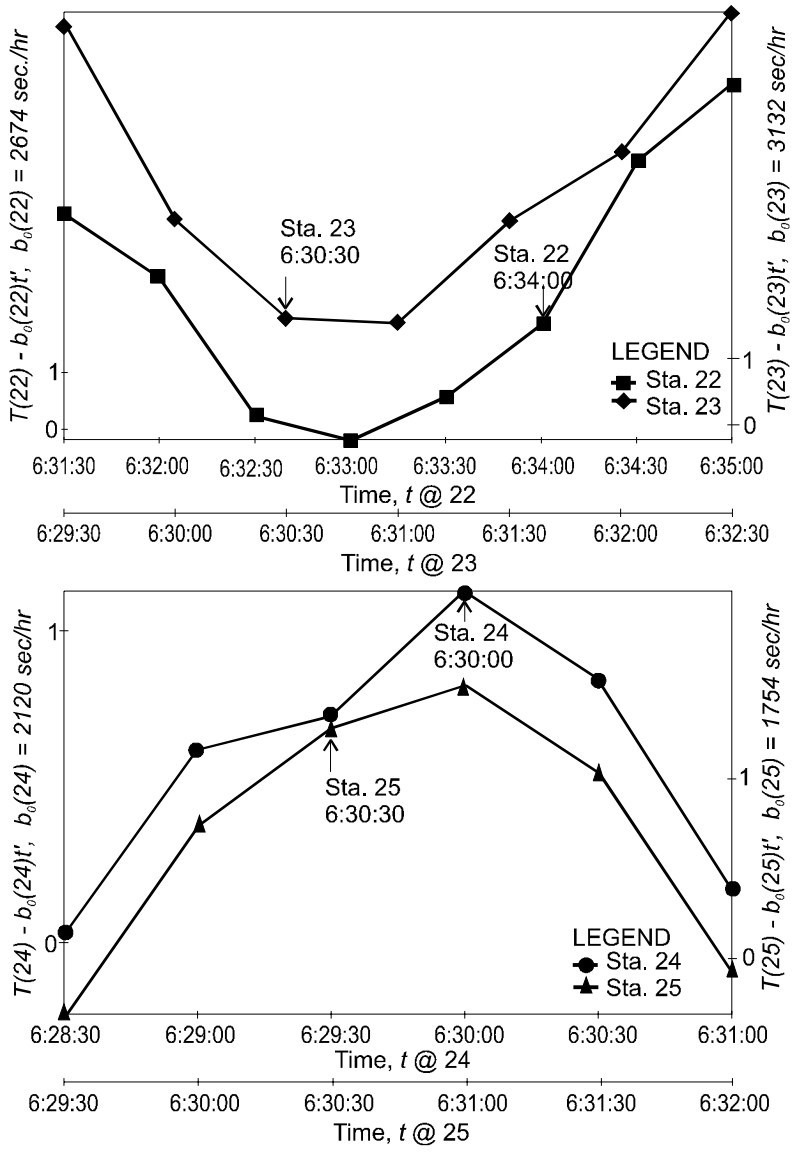

Figure 6-7. Re-scaled T-curves.

reductions previously revealed by the *N*-curves; the times marking the onset of these flow reductions are labeled on the lower portion of Figure 6-7. Conversely, the upper half of Figure 6-7 indicates that upstream detectors 23 and 22 each measured a rather abrupt increase in the occupancy rates. These occupancy changes coincided closely with the flow reductions previously identified at these detectors. The times marking the onsets of these upstream flow reductions are labeled in the top half of the figure.

The interpretation. The traffic patterns described above reveal that a forward-moving interface signaling lower flow and occupancy,[2] along with a backward-moving queue, emanated from between detectors 23 and 24. This confirms that a bottleneck was activated somewhere in the intervening segment.

Repeated observations. As part of a study on freeway capacity (Cassidy and Bertini, 1999), traffic conditions on the freeway segment in Figure 6-5 were examined using data collected on several weekday mornings. During each of these mornings, the bottleneck formed between detectors 23 and 24. The reason(s) that the bottleneck occurred about a kilometer or more downstream of the merge is a subject of ongoing research. For now, it suffices to note that a bottleneck's location is not always obvious. This, in turn, underscores the value of the diagnostics illustrated above.

The bottleneck's persistence. Knowing the duration that a bottleneck remains active is important for certain tasks such as estimating the bottleneck's capacity or predicting the evolution of its queue. (These topics are covered later in this section and in section 6.4). Toward determining a bottleneck's persistence, one can construct transformed *N*-curves that are similar to those in Figure 6-6, but that have been constructed using counts taken over an extended time period; i.e., one that spans the entire rush. By constructing them over a prolonged period, the displacements in the curves reveal the persistence of upstream queuing. Some examples of this are provided in Cassidy and Bertini (1999).[3]

It is likewise advantageous to construct re-scaled curves of *N* and of *T*, again for extended durations spanning the rush, using measurements taken downstream of the bottleneck. (In the context of the present example, curves could be constructed from measurements at detectors 24 and/or 25). One can examine these *N* and *T*

[2] One would expect to view this type of forward-moving interface following a sudden restriction in the flow upstream. This characteristic has been consistently observed to accompany the onset of upstream queueing, despite the notable absences of any traffic accidents or other exogenous causes of flow reduction (Cassidy and Bertini, 1999). Why queue formations give rise to this traffic characteristic is currently under investigation.

[3] Comparing cumulative curves that have been constructed over long time periods requires accurate measurements since errors can accumulate over time.

collectively to check for the arrival of any queue that may have spilled-over from further downstream and restricted flow through the bottleneck of interest. Such an arrival is marked by a reduction in the N accompanied by an increase in the T, as shown previously. Long-run N- and T-curves constructed at detector 25 are shown as part of the discussion presented next.

Estimating Bottleneck 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

$$q_{max} \equiv \lim_{T \to \infty} \left(\frac{N_{max}}{T} \right), \tag{6.20}$$

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 $\hat{q}_{max}$ (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, $E(\hat{q}_{max}) \approx q_{max}$. For this reason, one would collect samples (counts) immediately downstream of an active bottleneck so as to measure vehicles discharging at a maximum rate. The amount that $\hat{q}_{max}$ 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

$$\hat{q}_{max} = \sum_{m=1}^{M} n_m / (M \cdot \tau), \tag{6.21}$$

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

$$variance(\hat{q}_{max}) = \frac{1}{\tau^2} \left[\frac{variance(n)}{M} \right] = \frac{1}{T} \left[\frac{variance(n)}{\tau} \right] \tag{6.22}$$

since q_{max} is a linear function of the independent n_m and the (finite) observation period T is the denominator in (6.21).

The bracketed term $variance(n)/\tau$ 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 $E(n)/\tau = q_{max}$, one obtains

$$variance(\hat{q}_{max}) = \frac{\gamma}{T}, \tag{6.23}$$

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 (6.23) and then multiplying both sides of the resulting expression by q_{max}, one arrives at

$$q_{\max} \cdot T = \frac{\gamma}{\varepsilon^2}, \tag{6.24}$$

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 $\varepsilon = 0.05$ 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.

A different definition of capacity. The *Highway Capacity Manual* (TRB, 1994), a widely circulated guidebook, offers a definition of capacity different from the one above. The *Manual* recommends taking as an estimate of q_{max} the highest flow measured over some interval, usually 15 minutes, during periods when "*sufficient demand exists*." The *Manual* is not specific about what constitutes the existence of "sufficient demand." This omission may be intentional since there appears to be a lack of consensus as to whether capacity is a bottleneck's long-run queue discharge rate or the higher flow sometimes reported to occur prior to upstream queueing (Agyemang-Duah and Hall, 1991; Banks, 1990, 1991; Cassidy and Bertini, 1999). A rationale for treating capacity as the former, and an illustration of a deficiency in the *Highway Capacity Manual's* recommendation, are both offered later in this section. For now, it suffices to note that (6.24) yields the sample size for estimating capacity to within a specified error given the traffic features (i.e., the γ) at the bottleneck of interest. This has obvious appeal as compared with taking samples over some interval of arbitrary duration, such as 15 minutes.

Also of note, the *Highway Capacity Manual* provides default values for estimating a bottleneck's capacity without collecting samples in the field. These default values may be useful for certain long-range planning applications when coarse estimates suffice. They should only be used, however, when suitable sample counts cannot be collected.

An illustration of freeway capacity. Figure 6-8 presents re-scaled *N*- and *T*-curves that were constructed for a period spanning the morning rush. These measurements were taken downstream of the bottleneck (at detector 25) which was found on this day to be active from 6:30:30 a.m. to 7:54:00 a.m., as labeled on the figure. During this period of nearly 90 minutes, the *N* and the *T* display similar features, indicating that the measurements were not influenced by traffic conditions (i.e., queues) from further downstream.

The *N*-curve reveals that the observed pattern of flow can be described as sequences of sustained surges followed by reductions. This pattern is highlighted in Figure 6-8 by means of linear approximations superimposed on the *N* and by labels designating the flows (in units of vehicles per hour) for each period marked by quasi-linear arrivals. Despite these variations, the queue discharge flows exhibit a constant long-run trend, shown by the dashed line in the figure. While the bottleneck was active, the *N* never deviated from this trend by more than about 50 vehicles. By constructing the *N*-curve with a smaller background flow reduction and/or by plotting it over a longer time, curve portions measured during the period of queue discharge would appear to have a constant slope. Thus, the bottleneck's queue discharge rate may be described as being nearly constant over the rush.

One could determine with (6.24) a suitable sample size for estimating the average discharge rate. In this case, it would be advantageous to estimate the γ with a presample that captures both the surges and the reductions in the discharge flow; e.g. the $\{n_m\}$ could be periodically sampled over the entire period marked by the discharging queue. In general, such a sampling scheme would be feasible. If, for example, measurements were taken (automatically) using loop detectors, counts would usually be available for the entire day, including the rush. If instead, one incurred the expense of collecting counts by deploying human observers roadside, it would make sense to extend the data collection period so as to include the entire rush.

In light of the above, it would make even more sense to treat the queue discharge flow observed over an entire rush as an estimate of the long-run average. Eq. 6.24 could be rearranged to obtain ε, the precision of this estimate.

Treating the long-run discharge rate as capacity. Findings indicate that a bottleneck's average discharge rate, when measured over the rush, is reproducible from day to day (Cassidy and Bertini, 1999). Given this reproducibility, and in light of its near-constancy over the rush, it seems reasonable to take the estimated long-run average discharge rate as the bottleneck capacity. In the present example, the

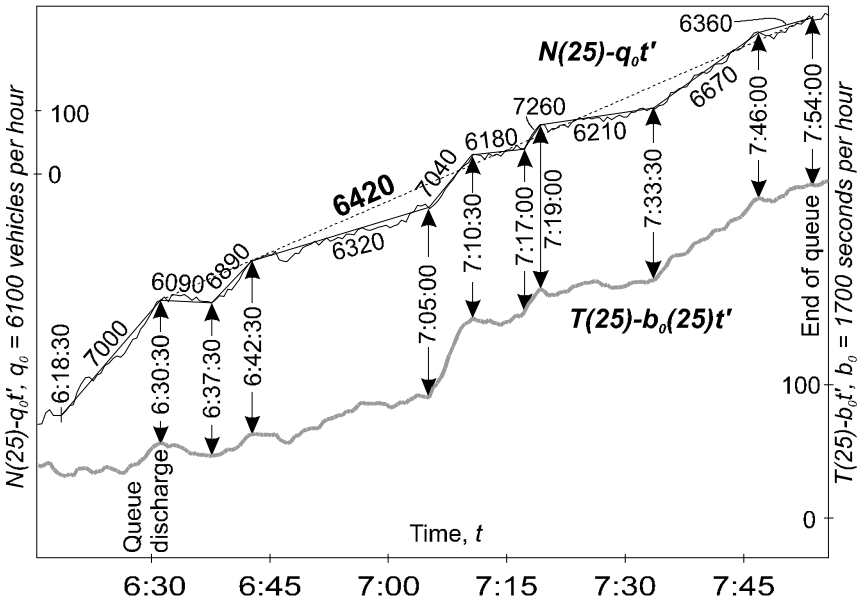

Figure 6-8. Re-scaled *N*- and *T*-curves downstream of the bottleneck.

estimated capacity is therefore 6,420 vehicles per hour (vph), as annotated on Figure 6-8.

Revisiting the Highway Capacity Manual's recommendation. If one takes as capacity the maximum flow observed over an interval of about 15 minutes (TRB, 1994), then, for the present example, the estimate is higher than 6,420 vph. This is because a flow of 7,000 vph prevailed for 12 minutes before the bottleneck's activation; this very high flow is labeled in Figure 6-8. Although some studies have reported that bottlenecks can support very high flows prior to their activation, these high flows have typically been observed only for time periods that are short relative to the rush. Figure 6-8 also shows that queue discharge rates comparable to 7,000 vph occasionally arose, but again, only for brief durations. Notably, these periods of very high flow are not only short-lived, their (short) durations vary from one day to the next (Cassidy and Bertini, 1999). Thus, these very high flows appear to be unstable and whether they can be prolonged through control measures such as on-ramp metering remains a question of active research. In light of this, the higher estimate of capacity (i.e., the one that follows from the *Highway Capacity Manual's* recommendation) may be unduly optimistic, even misleading.

6.3 Intersection Capacity and Vehicle Delay

This section describes techniques for estimating the capacity of highway intersections controlled by either traffic signals or stop signs, as these are often bottlenecks. Methods of estimating the vehicle delays at these facilities are likewise discussed, since delay minimization is a commonly used objective when developing intersection control schemes. This section does not specifically address such control schemes, however, as these are covered in Chapter 7 of the handbook.

Signalized Intersections

At a busy intersection, a traffic signal periodically interrupts vehicle movements to serve traffic in conflicting directions. Green times are extended so that consecutive vehicles in a queue may discharge through the intersection at a high rate, termed the saturation flow. It is by serving vehicle movements in this batched manner that a signal can increase the rates by which (conflicting) traffic streams traverse the intersection.

Much like the queue discharge rates through freeway bottlenecks, an intersection's saturation flows are affected by its geometric features and by certain attributes of its traffic streams such as the percentage of trucks. Moreover, saturation flows often vary with the type of traffic movement (i.e., turning or through-moving) served. Intuitively, a traffic movement's capacity is the product of its saturation flow and the proportion of "green time" available for this discharge. Methods of capacity estimation are described below.

Saturation flow and capacity. If by the end of the signal's red time, a traffic movement exhibits a sufficiently long queue (perhaps 6 or more vehicles in each lane), one can estimate its saturation flow by sampling the times consecutive vehicles pass a fixed point near the intersection, such as the stop bar, and plotting these departures cumulatively. Figure 6-9 shows a hypothetical cumulative count curve for the vehicles observed (e.g. in a single lane) during a time period greater than one signal cycle length, C. The step-wise curve has a strictly horizontal portion beginning some time near the end of the green indication. This period of "zero flow" extends until some time after the initiation of the green because the queued vehicles do not begin discharging at the instant of the green's initiation (Webster, 1958).

By collecting the departure times over K cycles and setting time t equal to zero at the initiation of each green, the arithmetic average of the cumulative number of vehicles to enter the intersection by t, $\overline{N}(t)$, is $\frac{1}{K}\sum_{k=1}^{K} N_k(t)$, where $N_k(t)$ is the cumulative number entering the intersection by t during the kth cycle. For a sufficiently large K, the cumulative curve is smooth, as exemplified in Figure 6-10. The slope of this smooth curve rises gradually from zero at $t = 0$ to a maximum of s,

the saturation flow. This maximum slope eventually transitions to a smaller value (equal to the average arrival rate, $\overline{q}$) because the queue vanished during the green.

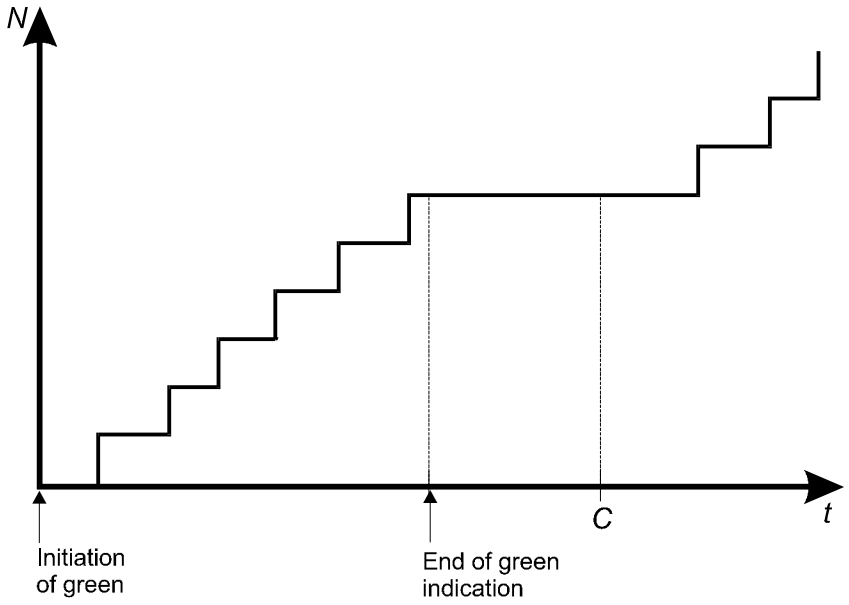

Figure 6-9. Typical N-curve at a traffic signal (Newell, 1989).

The "effective" start of the green time can be identified by extrapolating the curve portion with slope *s* backwards in time until it intersects with the line $\overline{N} = 0$. In similar fashion, the end of this green is found (approximately) by extrapolating the linear curve portion with slope $\overline{q}$ forward in time until it intersects with an extrapolation of the horizontal curve portion, as shown in Figure 6-10. The effective green period, *G*, is the time available in each cycle for serving vehicles at rate *s*. Thus, the intersection's capacity to serve the traffic movement is $s \cdot (G/C)$.

One can usually assume that an $\overline{N}$-curve like the one described above is reproducible from day to day. Moreover, the precision of an estimate of *s* may be obtained using Eq. 6.24.

If drawn on "standard sized" paper, the $\overline{N}$-curve would not require any re-scaling, such as a background flow reduction. This is because the $\overline{N}$ and the *t* used to diagnose the intersection flows are small as compared with those used in diagnosing a freeway bottleneck.

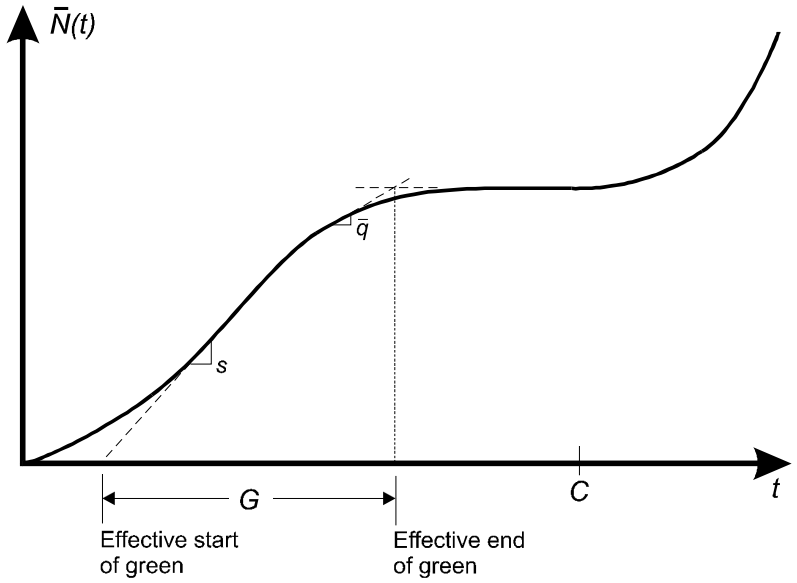

Figure 6-10. Average counts at a traffic signal, interpretation of effective green time (Newell, 1989).

Vehicle delay. If the traffic movement of interest is under-saturated (i.e., if the $\bar{q}$ does not exceed its capacity), an average departure curve like the one in Figure 6-10 can be used for measuring the vehicle delays. To this end, the vehicle arrival times may be measured upstream of any queueing caused by the traffic signal and a cumulative curve of the average arrivals per cycle can be constructed in the manner just described. One obtains a queueing diagram, and with it delay information, by horizontally translating this average arrival curve so that it is superimposed on the average departure curve for the period when the queue was not present.

Notably, the average arrival curve is merely a straight line (with slope equal to $\bar{q}$) if the vehicle arrivals are not influenced by any other traffic control device located further upstream. In these instances, the data needed to construct a queueing diagram can be collected by a single observer. A complete set of instructions for conducting such an experiment is provided by Pitstick (1990).

In many cases, one can predict intersection delays at some future time merely by altering the average arrival curve and/or the average departure curve as appropriate for the conditions (e.g. arrival rates, signal timing, etc.) that have been projected. The reader may again refer to Pitstick for discussion of this.

Stop-controlled Intersections

At an approach controlled by a stop-sign, the capacity to serve a certain traffic movement may be estimated by sampling the times that its queued vehicles enter the intersection. The issues here are much like those described in section 6.2, with the

notable difference that, for a traffic movement controlled by a stop sign, capacity is influenced by the conflicts created by vehicles on other approaches.

The effects of vehicular conflicts are especially complex when stop signs are used to control only the traffic approaching from the minor street. The capacities for these minor street movements depend upon the propensity of its drivers to utilize headways exhibited by vehicles entering from the major street; more precisely, these drivers are forced to use "gaps" in the major street's traffic streams(s). Complexities arise because this gap-acceptance behavior is influenced by a host of factors, including the geometry of the major street and the velocity of the vehicles traveling on it, the sight distance(s) available to drivers on the minor street, their intended maneuvers and their personalities (e.g. aggressive or timid).

Gap acceptance models and delay prediction. Models for predicting gap-acceptance behavior have been developed to serve as tools for planning purposes. Some of these assume that drivers are both homogeneous and consistent; i.e., presumably a gap larger than some critical value is invariably utilized by the motorist waiting at the stop bar and motorists always decline to use a gap that is smaller than this critical value (TRB, 1994). Other models assume that the critical value used for accepting or rejecting a gap varies across drivers (Cohen, et al., 1955; Solberg and Oppenlander, 1966; Miller, 1972; Daganzo, 1981). Some gap acceptance functions assume that a driver's critical value may change, for example, as she grows impatient while waiting at the stop bar (Mahmassani and Sheffi, 1981; Madanat, et al., 1994; Cassidy, et al., 1995).

To predict vehicle delays at stop-controlled intersections, gap acceptance functions have been incorporated into analytical queueing models or they have been used in computer simulations. As a practical matter, it is worth noting that intersection delays become substantial only when the traffic intensity (i.e., the ratio of arrival rate to capacity) reaches a value of about 0.8 (Webster, 1958). Such high ratios are seldom observed at stop-controlled intersections because of often-used warrants (MUTCD, 1988) that provide for the installation of a traffic signal even when traffic intensities are rather small. Moreover, all gap acceptance models are estimated through statistical means and the applicability of any one model is limited by the intersection characteristics used for its estimation.

6.4 Traffic Flow Theory

It was noted in the previous two sections that bottlenecks exhibit predictable features. In section 6.2, for example, the discharge flow from an active freeway bottleneck was shown to be nearly constant over the rush. It was further noted that freeway bottlenecks consistently arise at the same locations and that a bottleneck's average discharge rate is reproducible from day to day. The queue discharge rates at signal-controlled and stop-controlled intersections can also be characterized as reproducible, as noted in section 6.3.

In light of their predictable features, it seems reasonable to theorize about how traffic evolves upstream and downstream of bottlenecks. Such theories have been developed and they may be used to predict important attributes of the traffic stream, such as a queue's growth due to accidents or geometric restrictions, the flow patterns generated by traffic control measures, etc. These theories rely upon given sets of boundary conditions and these might entail, for example, the rate of trip generation at all origins along the highway, the routes of all trips and the vehicle trip times, where the latter are functions of the time-dependent flows along the highway.

This section begins with a description of a remarkably simple theory of highway traffic flow. Proposed by Newell (1993), this theory is a version of one originally developed by Lighthill and Whitham (1955) and by Richards (1956) whereby traffic is treated as a continuum; i.e., the models describe the collective or average motion of vehicles in a traffic stream. Newell's version is described as being a "simplification," partly because of the relation it assumes, in effect, between flow and vehicle trip time and an explanation of this is provided later in the section. The simple theory predicts the shapes of standard (i.e., untransformed) cumulative count curves at locations of interest along the roadway. With this as its framework, the theory exploits the advantages of cumulative curves already described in the previous three sections. Accordingly, the theory is presented here in a purely graphical way. To highlight the theory's intuitive attributes, it is described in the context of a simple scenario involving a single highway segment with a bottleneck of fixed capacity located somewhere further downstream. References are made to some of the empirical findings that support the theory. Extensions needed to apply the theory to more complex scenarios are likewise noted. Prior to concluding this section, some of the limitations of the simple theory (and of its predecessors) are mentioned and attention is briefly given to some other traffic theories that have been proposed in light of these limitations.

A Simple Theory of Traffic Evolution

The upper portion of Figure 6-11 presents some hypothetical trajectories passing measurement (e.g. detector) location x_1 and proceeding past downstream measurement location x_3. It is assumed here that the intermediate location x_2 has no device(s) for traffic measurement. Thus, the theory will be used to describe the time-dependent traffic conditions, or more specifically, to construct the cumulative count curve at this location. The reader will note that the theory could be used for constructing the cumulative curves at any intermediate locations of interest. To satisfy an essential boundary condition, the cumulative curve at x_1, $N(t, x_1)$, used for the present example was not affected by any queueing from downstream.

The trajectories drawn in Figure 6-11 describe one freely flowing traffic state, labeled *a*, and three different states in queued traffic, labeled *b* through *d*. All vehicles are assumed to exhibit identical headways, spacings and velocities within a given state. Thus, no vehicle overtaking occurs. Furthermore, all vehicles are assumed to travel at a free-flow velocity v_f whenever traffic is freely flowing. Thus, for the interval $t_2 - t_1$, the cumulative count curve at x_2, is obtained by constructing

the N-curve at x_1 and shifting it horizontally to the right by a vehicle's free-flow trip time from x_1 to x_2, $1/v_f \cdot (x_2 - x_1)$. The curve labeled I was constructed by translating $N(t, x_1)$ in this manner. (Step-wise curves are shown in Figure 6-11 to make more obvious the relation between the trajectories and the cumulative curves).

In queued state b, the lead vehicle, labeled 0, decelerated from its previous (free-flow) velocity. It was stopped in state c and it accelerated upon entering

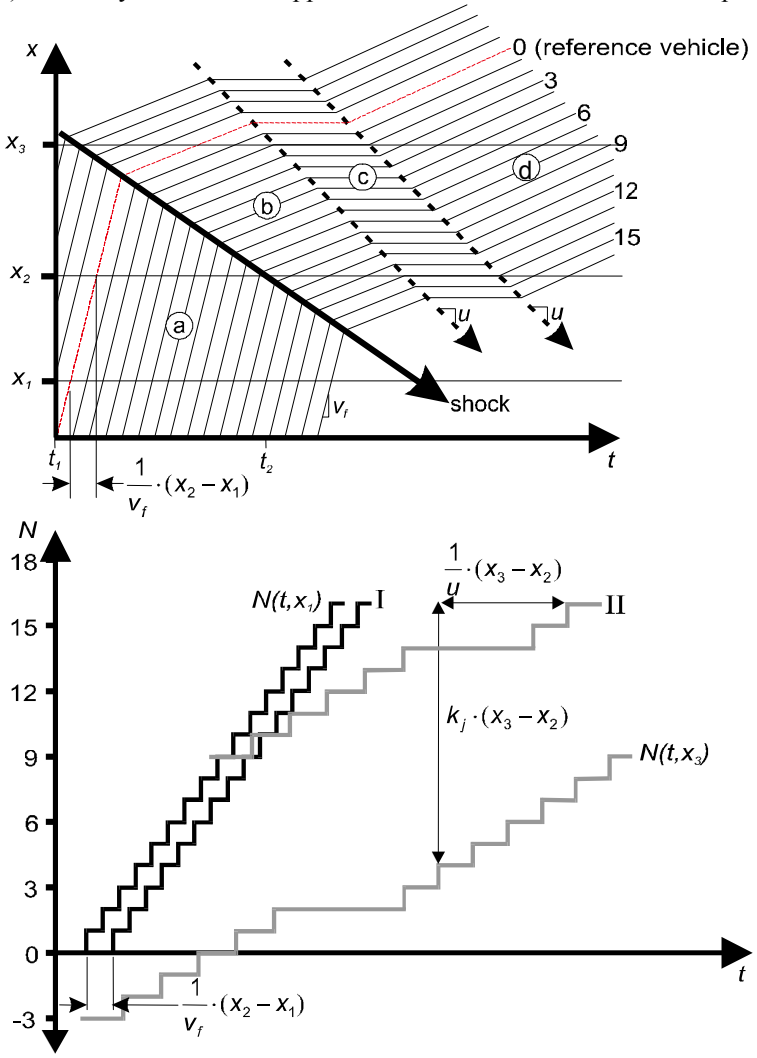

Figure 6-11. Simple example of traffic evolution.

state *d*. Of note, the theory assumes that vehicle accelerations (and decelerations) occur instantaneously and thus the theory can only hold over dimensions of time and space that are large relative to the separations between vehicles.

All vehicles of higher arrival number behave precisely as vehicle *0*. In queued traffic, an *i*th vehicle's trajectory is assumed to adopt the features of the *i-1*th trajectory following some fixed time lag; the time lag is the same for all queued states. This, along with the assumption that vehicles exhibit uniform spacings within a given state, means that the interface between any two queued states propagates at a fixed speed u, as shown by the dashed lines in the upper portion of Figure 6-11. Moreover, in queued traffic, the *i*th trajectory can be constructed by shifting the *i-1*th trajectory forward by the fixed time lag and downward by a fixed spacing. Study of the time-space diagram in Figure 6-11 reveals that this spacing is the one adopted by vehicles that have come to a complete stop, the so-called jam-density spacing. This is true even in the absence of a "jammed" traffic state like state *c*. In fact, state *c* was included in Figure 6-11 merely to aid the reader in verifying that the appropriate downward translation is always the jam density spacing.

It follows that in queued traffic, the curve at x_2 is obtained by shifting the *N*-curve at x_3, $N(t, x_3)$, to the right by a distance equal to the trip time of an interface between these two locations, $1/u \cdot (x_3 - x_2)$, and upward by the number of vehicles that pass through the interface during this time. It should be clear that the latter is the jam-density storage, $k_j \cdot (x_3 - x_2)$, where k_j , the jam density, is the maximum density the road segment can accommodate.

These horizontal and vertical curve translations produced the curve labeled *II* in Figure 6-11. By referring to the time-space diagram in this figure, the reader may verify that shifting $N(t, x_3)$ as described above produces the *N*-curve that would have been measured for queued traffic conditions at x_2, had measurement devices existed there.

The curve translated forward from x_1, labeled *I*, intersects the curve translated from downstream x_3, labeled *II*. Notably, this intersection occurs at time t_2, the time when the back of the queue (i.e., state *b*) arrived to location x_2. Time t_2 is thus said to mark the arrival of a *shock* at location x_2, where, in this simple theory, a shock is an interface between queued and freely flowing traffic.[4] The lower envelope of curve *I* and curve *II* in Figure 6-11 is the resulting *N*-curve at x_2. It is intuitive that flow is constrained at x_2 following the shock's arrival and the two translated curves intersect at this arrival time because vehicles are conserved across the shock's path.

A shock may exhibit a multitude of possible speeds, positive or negative, as dictated by the traffic conditions on its upstream and downstream sides. Moreover, a shock's speed changes when it intersects interfaces or other shocks. Notably, the use of *N*-curves as described above does not require one to trace the paths of shocks and

[4] In the theory developed by Lighthill and Whitham and by Richards, shocks arise when backward-moving interfaces collide. These collisions do not occur in the simplified theory since all backward interfaces are presumed to travel at the same speed u.

interfaces over time and space. Such (tedious) analyses are required in the Lighthill and Whitham and Richards (LWR) versions.

An assumed bivariate relation. Implicit in the simple theory and LWR is a key postulate that there exists some relation between traffic properties, such as flow and vehicle trip time, that may vary with location along the highway, but not with time. These relations are purely empirical; i.e., they are obtained through measurement. Not only would they depend upon the highway geometry, these relations would also be affected by environmental factors, such as weather conditions, as well as by attributes of the traffic stream, such as its proportion of large trucks, the tendencies of its drivers, etc. Although the bivariate relation between any number of traffic properties may be used as a boundary condition for continuum models, it is customary to use a relation between flow, q, and density, k. The assumptions of the simple theory just described imply a q-k relation that is triangular in form, like the one shown in Figure 6-12, and the reasons for this are explained below.

To begin, the right branch of the relation describes conditions in queued traffic, whereby k increases with decreasing q. From Eq. 6.13, the speed of an interface between stationary traffic states, u, is $\Delta q/\Delta k$. That the relation's right branch is linear means that interfaces between queued states presumably propagate at a single speed, a previously noted assumption of the simple theory.

The left branch of the triangular relation describes freely flowing traffic. As noted in section 1, the average vehicle velocity is the ratio of flow to density. Thus, the simple theory assigns to the left branch a slope of v_f, the presumed velocity of all freely flowing vehicles. That this slope is constant also implies that changing (q, k) states move forward with the vehicles in freely flowing traffic.

It was previously noted that, in the simple theory, a shock separates freely flowing and queued traffic. The shock's speed is therefore the slope of the chord connecting the (q, k) states as they lie on each side of the triangular relation.

The existence of well-defined bivariate relations is an assumption widely adopted in traffic and transportation engineering. In addition to their role in continuum theories of traffic flow, these relations are commonly used in highway design and in transportation planning. In fact, additional discussion on these presumed relations may be found in virtually any traffic engineering text or handbook, including (TRB, 1994). In particular, chapter 4 of Daganzo (1997) contains a thorough introduction to the subject.

Some empirical evidence. Despite their role in continuum theories like the simple one described above, the assumption that bivariate relations are independent of time is known to be incorrect. Interfaces in the traffic stream exhibit characteristic widths where traffic is not stationary because vehicles are adjusting from one state to another. When collected in the presence of these nonstationary regions, bivariate data do not give rise to well-defined relations. To the contrary, plots of these data are widely scattered (Hall, et al., 1992).

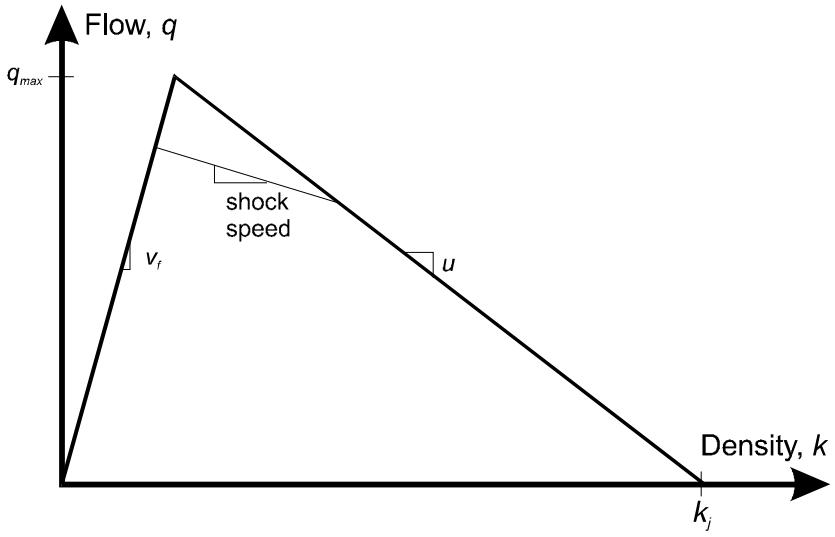

Figure 6-12. Triangular flow-density relation.

On the other hand, one might "*reasonably expect drivers to do the same on average under the same average conditions*" (Daganzo, 1997, p 80). It therefore seems reasonable to postulate that well-defined bivariate relations exist for stationary traffic. Indeed, there is the empirical evidence supporting this postulate. One study found that the average values of flows and occupancies taken only from nearly stationary traffic produced well-defined relations that appeared to be reproducible from day to day (Cassidy, 1998). The research showed that (freeway) traffic is stationary in the absence of any interfaces and that one may describe these conditions on a highway segment using some bivariate relation. It follows that an interface must propagate as specified by the highway segment's bivariate relation, even though the relation does not describe traffic behavior in the immediate vicinity of the interface.

In another study, re-scaled cumulative count curves were used to examine numerous backward-moving interfaces in individual travel lanes (Windover, 1998). These curves were constructed in series using counts taken (by loop detectors) on a 2-km freeway segment upstream of an active bottleneck. They were plotted for time periods long enough so as to display numerous interfaces and these interfaces separated a variety of queued (*q, k*) traffic states. The curves were observed to have similar forms; i.e., slope changes on the downstream curve later appeared on the upstream curves, indicating that drivers responded in similar ways to changes in traffic conditions as per LWR and Newell's simple theory. Furthermore, each (entire) curve could be nearly superimposed upon its neighbor following a vertical and a horizontal translation. This latter observation indicates that between any two

of the measurement locations, all interfaces passed through nearly the same number of vehicles and had nearly the same trip times. The finding indicates that the adoption of triangular *q-k* relations like the one in Figure 6-12 may be a reasonable approximation for describing traffic.

Applying the simple theory to complex scenarios. The simple theory was described above using a highway segment of fixed geometry and a single bottleneck downstream. Newell describes how the theory can be applied to more complicated roadways by using suitable boundary conditions. In addition to those required for the previous analysis, these boundry conditions include the differences between the cumulative number of vehicles that enter and exit the roadway by time *t* at each junction; i.e., the net cumulative count.[5] One would also specify (possibly different) triangular *q-k* relations for each section of highway.

As regards the net cumulative counts, a complication arises in estimating the exit counts at each junction. Even if one knows the routes taken by vehicles and the times they enter the roadway, the times they actually exit will be dictated by the prevailing traffic conditions whenever queueing occurs upstream. This complication was originally recognized by Vaughn, Hurdle and Hauer (1984) and also by Vaughn and Hurdle (1992). Newell (1993, part *III*) provided a solution by specifying that a vehicle's trip time between successive junctions is independent of its origin and destination and that at all junctions, one must evaluate the component cumulative curves for each destination. The reader may refer to the source for a complete description of this.

On a related note, a diverge formed by an off-ramp can cause queueing in at least two ways: 1) a queue from the off-ramp spills over and blocks traffic; and 2) the off-ramp is not queued but an increase in the flow of vehicles wishing to exit creates a queue (on the freeway or highway) when the off-ramp reaches capacity (Daganzo et al., 1997). A mathematical theory of the diverge can be found in sections 3.3 and 4 of Daganzo (1995) and a preliminary theory for wide freeways where traffic may sort itself by lane depending upon destination can be found in Daganzo (1997*a*).

An additional complication arises in determining the net cumulative count at some on-ramp after a queue has propagated beyond it. This is because the rate by which vehicles merge onto a congested roadway requires some observation or some intuition. Studies by the California Department of Transportation (Newman, 1986), for example, found that when queues are present both on an entrance ramp and on the freeway, merging vehicles share available roadway space with vehicles in the adjacent freeway lane on a one-to-one basis, creating the so-called "zipper effect." Daganzo (1997a) offers a simple theory to predict both the ramp and the freeway (output) flows when two merging traffic streams compete for the capacity available downstream.

Finally, a computer program has been developed to aid in applying Newell's simple theory to complex highway sections. As of the date of this publication, the

[5] If entrance and exit ramps are closely spaced, their separations may be ignored.

program is available on the World Wide Web through the civil engineering department at the Georgia Institute of Technology.

Some deficiencies and some alternatives. The simple theory, as well as the LWR theory, are known to have deficiencies. The most noteworthy of these, along with some of the other theories developed to address these deficiencies, are described below. The coverage here is admittedly light. The intent is merely to bring toward closure the discussion of traffic flow theory and to provide the reader with references to some of the other models common in the literature.

Traffic instabilities. As apparent from the previous discussion, the simple theory does not predict the "stop and go" instabilities often observed in queues. Nor are instabilities described by the LWR theory. Models that qualitatively match many (although not all) of the features observed in unstable traffic do exist. One of the earliest of these (Newell, 1962) belongs to a class of models that are, in effect, extensions to the LWR theory and its simplified version. This class of models share a number of similarities with the latter continuum models. In all of these models, for example, a platoon is described by state (q, k) on the flow-density plane. Likewise, each of these models assume that all vehicles in a platoon always respond in the same way to a sustained change in the lead car's velocity. However, the class of models which we describe as being extensions are specified by defining two families of q-k curves; one family describes all possible evolutions of decelerating platoons and the other corresponds to accelerating traffic (Daganzo, et al., 1997). Models of this type are thus compatible with a well-known finding of Edie (1965) that decelerating platoons adopt a set of states on the q-k plane that are consistently different from accelerating ones.

Efforts to explain traffic instabilities also led to the development of models to describe a driver's response to the changing trajectory of a lead vehicle. These so-called car-following or microscopic traffic models predict how a driver adjusts her vehicle's velocity because of a stimulus, such as the (time-dependent) spacing between the subject vehicle and its leader (Chandler, et al., 1958; Herman, et al., 1959; Gazis, et al. ,1959; Herman and Potts, 1961). To calibrate and test these models, considerable data collection took place using instrumented cars on test tracks. This line of research is notable in that the objective was to describe the behavior of individual drivers and the interested reader can refer to May (1990) for discussion on some of the early developments in this area.

To their credit, the car-following models do predict the occurrences of instabilities. However, there has been little success in matching the oscillatory behavior observed in real traffic with the predictions from these car-following models. Perhaps this is because the theories assume that drivers apply control continuously and that a driver responds only to the motion of the car immediately downstream and not, for example, to cues collectively displayed by the (possibly many) downstream vehicles visible to the driver. Driver behavior is probably more complex than this.

Indeed, the complexities of driver behavior likely explain why instabilities exhibit periods of oscillation and growth that are site specific. For example, a bottleneck studied in New York's Holland Tunnel displayed regular oscillation periods of about 2 minutes (Edie and Foote, 1961), while observations upstream of other bottlenecks have revealed more sporadic characteristics (Kerner and Rehborn, 1997, 1996, 1996*a*, Smilowitz et al., 1998). This is not surprising given that the vehicular interactions are different for different types of bottlenecks (e.g. merges, diverges, lane reductions, etc.) and a thorough understanding of each type of bottleneck will likely come only by studying them individually.

What may be most important here, however, are the details of traffic instabilities that are understood at the present time. Namely, that instabilities occur well upstream of bottlenecks and that they do not appear to affect a bottleneck's discharge flow.[6] The detailed behavior of queued traffic therefore has little effect on the delay caused by a bottleneck. Furthermore, to predict time-dependent queue lengths, (e.g. due to control actions) it might suffice to predict the general, coarse shapes of cumulative count curves without attempting to predict the occurrences of wiggles that characterize instabilities. This is precisely the objective behind Newell's simple theory.

Theories from other scientific fields. In efforts to improve the LWR models, researchers have borrowed theories from other fields of scientific inquiry. Discussion of methods adopted from other fields are not the emphasis of this handbook. Nonetheless, a few of these adaptations deserve mention, albeit brief, because they are prominent in the literature, they are used in practice and they are not without shortcomings of their own.

For example, Navier-Stokes-like equations commonly used in fluid mechanics have been proposed for describing traffic instabilities (Kerner, et al., 1995). Researchers have likewise advocated other second-order partial differential equations similar to those used for fluid approximations to explain the motion of vehicles passing through shocks (Payne, 1971; Kühne and Beckschulite, 1993); the reader will recall that the simple theory and its predecessors assume that vehicles change velocities instantaneously and this is a very coarse approximation. These second-order models can generate unrealistic predictions of traffic evolution because these models borrow features of materials flow that are unreasonable for describing highway traffic. For example, del Castillo et al. (1993) have noted that these models predict that interfaces may overtake particles (e.g. fluid molecules) and that when applied to highway traffic, this implies that drivers collectively respond to stimuli from behind. This would not seem to be a reasonable depiction of the driving

[6] Contrary to claims made in some of the literature (Kerner, et al., 1995; Kerner and Rehborn, 1997), this author has seen to date no conclusive evidence suggesting that instability phenomena may cause freely flowing traffic to break-down and form queues spontaneously. The interested reader may refer to Daganzo et al. (1999) for more discussion on these issues.

process. Daganzo (1995a) offers extended discussion on the pitfalls in applying second-order fluid approximations to highway traffic.

Lastly, models based upon the kinetic theory of gases have been proposed for describing vehicular interactions in light traffic (Phillips, 1977; Prigogine, 1961; Prigogine and Herman, 1971). Continuum theories are deficient for these conditions because they do not describe the variations in velocities across vehicles and thus, they do not predict the overtaking and the natural spreading of platoons that occur in low densities. The kinetic models of traffic flow were intended to improve the LWR theory by considering the distribution of vehicle velocities at each point in time and space. However, these models were derived from the integration of molecular properties, such as positions, collisions and velocities, that do not accurately describe highway traffic. Daganzo (1995a), for example, has noted that these models assume that a distribution of desired vehicle velocities "*can be defined exogenously at every point in time and space independent of the drivers who happen to be there*." In so doing, these models fail to recognize that individual drivers have personalities (e.g. aggressive and timid) which they retain with their motion (Cassidy and Windover, 1998). The reader may refer to Newell (1995) for further discussion on these issues. As an aside, a theory of very light traffic with weak overtaking interactions is fairly complete and its description is likewise found in Newell (1995).

Some Final Comments

Although section 6.4 has made reference to a number of theories for describing highway traffic, it has emphasized a simple continuum model proposed by Newell. In the interest of completeness, we note that Newell described his recipe using a coordinate system whereby the time at each location along the roadway is measured from the passage of a freely-flowing reference vehicle (e.g. labeled $N = 0$). In effect, this so-called "moving time coordinate system" horizontally shifts the cumulative count curves so that neighboring curves display vehicle delays and excess accumulations. This is analogous to the horizontal shifts that were applied to count curves in section 6.2 and the reader may refer to the original source (Newell, 1993) for more details on the use of moving time.

Far more important than these details, however, is the question of reliability; i.e., the adequacy of Newell's simple theory for predicting traffic evolution remains an active research question. The need for changing or refining the model may become apparent, for example, as ongoing empirical studies reveal more about roadway bottlenecks and the features of the queues they create. What is noteworthy about Newell's recipe, however, is its exploitation of cumulative count curves. Such a framework can be used to predict virtually any traffic feature likely to be of interest, including vehicle delays, the spatial extent of queueing, etc. Thus, it would seem that any models of highway traffic flow developed in the future, or any future refinements to existing models, should make use of these cumulative curves.

6.5 References

Agyemang-Duah K. and Hall F.L. (1991) Some issues regarding the numerical value of capacity. *Proc. Int. Symp. on Highway Capacity*, pp. 1-15, A.A. Balkema, Germany.

Banks J.H. (1990) Flow processes at a freeway bottleneck. *Transpn Res. Rec.* ***1287***, 20-28.

Banks J.H. (1991) Two-capacity phenomenon at freeway bottlenecks: a basis for ramp metering? *Transpn Res. Rec.* ***1320***, 83-90.

del Castillo J.M., Pintado P. and Benitez F.G. (1993) A formulation for the reaction time of traffic flow models. *Proc. Int. Symp. on Transportation and Traffic Theory*, (C.F. Daganzo, ed.), pp. 387-405, Elsevier, New York.

Cassidy M.J. (1998) Bivariate relations in nearly stationary highway traffic. *Transpn Res.* ***32B***, 49-59.

Cassidy M.J. and Bertini R.L. (1999) Some traffic features at freeway bottlenecks. *Transpn Res.* ***33B***, 25-42.

Cassidy M.J. and Coifman B. (1997) Relation among average speed, flow and density and the analogous relation between density and occupancy. *Transpn Res. Rec.* ***1591***, 1-6.

Cassidy M.J., Madanat S.M. and Wang M.H. (1995) Unsignalized intersection capacity and level of service: revisiting critical gap. *Transpn Res. Rec.* ***1484***, 16-23.

Cassidy M.J. and Windover J.R. (1995) Methodology for assessing dynamics of freeway traffic flow. *Transpn Res. Rec.* ***1484***, 73-79.

Cassidy M.J. and Windover J.R. (1998) Driver memory: motorist selection and retention of individualized headways in highway traffic. *Transpn Res.* ***32A***, 129-137.

Chandler R.E., Herman R. and Montroll E.W. (1958) Traffic dynamics: studies in car-following. *Opns. Res.* ***6***, 165-184.

Cohen, E., Dearnaley J. and Hansel C. (1955) The risk taken in crossing a road. *Opns. Res. Qrtly,* ***6***, 120-128.

Daganzo C.F. (1981) Estimation of gap acceptance parameters within and across the population from direct roadside observation. *Transpn Res.* ***15B***, 1-15.

Daganzo C.F. (1995) The cell transmission model. *II*: Network traffic. *Transpn Res.* ***29B***, 79-93.

Daganzo C.F. (1995*a*) Requiem for second-order fluid approximations of traffic flow. *Transpn Res.* ***29B***, 277-286.

Daganzo C.F. (1997) *Fundamentals of transportation and traffic operations*. Elsevier, New York.

Daganzo C.F. (1997*a*) The nature of freeway gridlock and how to prevent it. *Proc. Int. Symp. on Transportation and Traffic Theory*, (J.B. Lesort, ed.), pp. 629-646, Pergamon, Tarrytown.

Daganzo C.F., Cassidy M.J. and Bertini R.L. (1999) Possible explanations of phase transitions in highway traffic. *Transpn Res.* ***33A***, 365-379.

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

Edie L.C. (1974) *Traffic Science* (D.C. Gazis, ed.), pp. 8-20, Wiley, New York.

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

Edie L.C. and Foote R.S. (1961) Experiments on single-lane flow in tunnels. *Proc. Int. Symp. on the Theory of Traffic Flow* (R. Herman, ed.), pp.175-192, Elsevier, New York.

Gazis D.C., Herman R. and Potts R.B. (1959) Car-following theory of steady state flow. *Opns. Res.* ***7***, 499-505.

Hall F.L., Hurdle, V.F. and Banks, J.H. (1992) A synthesis of recent work on the nature of speed-flow and flow-occupancy (or density) relationships on freeways. *Transpn Res. Rec.* ***1365***, 12-18.

Herman R., Montroll E.W., Potts R.B. and Rothery R. (1959) Traffic dynamics: analysis of stability in car-following. *Opns. Res.* ***7***, 86-106.

Herman R. and Potts R.B. (1961) Single-lane traffic theory and experiment. *Proc. Int. Symp on the Theory of Traffic Flow* (R. Herman, ed.), pp 120-146, Elsevier, New York.

Kerner B.S., Konhauser P. and Schilke M. (1995) Deterministic spontaneous appearance of traffic jams in slightly inhomogeneous traffic flow. *Phys. Rev. E* ***51***, 6243-6246.

Kerner B.S. and Rehborn H. (1996) Experimental properties of complexity in traffic flow. *Phys. Rev. E* ***53***, R4275-R4278.

Kerner B.S. and Rehborn H. (1996*a*) Experimental features and characteristics of traffic jams. *Phys. Rev. E* ***53***, R1297-R1300.

Kerner B.S. and Rehborn H. (1997) Experimental properties of phase transitions in traffic flow. *Phys. Rev. Let.* ***79***, 4030-4033.

Kühne R.D. and Beckschulte R. (1993) Non-linearity stochastics in unstable traffic flow. *Proc. Int. Symp. on Transportation and Traffic Theory* (C.F. Daganzo, ed.), pp. 367-386, Elsevier, New York.

Lighthill M.J. and Whitham G.B. (1955) On kinematic waves. *I*: Flood movement in long rivers. *II*: A theory of traffic flow on long crowded roads. *Proc. Royal Soc.* ***A229***, 281-345.

Lin W.F. and Daganzo C.F. (1997) A simple detection scheme for delay-inducing freeway incidents. *Transpn Res.* ***31A***, 141-155.

Madanat S.M. Cassidy, M.J. and Wang M.H. (1994) A probabilistic model of queueing delay at stop-controlled intersection approaches. *J. Transpn Eng.* ***120***, 21-36.

Mahmassani H. and Sheffi Y. (1981) Using gap sequences to estimate gap acceptance functions. *Transpn Res.* ***15B***, 243-248.

Makagami Y., Newell G.F. and Rothery R. (1971) Three-dimensional representation of traffic flow. *Transpn Sci.* ***5***, 302-313.

May A.D. (1990) *Traffic flow fundamentals*. Prentice Hall, New Jersey.

Miller A. (1972) Nine estimators of gap acceptance parameters. *Proc. Int. Symp. on the Theory of Traffic Flow and Transportation*, (G.F. Newell, ed.), pp. 215-235, Elsevier, New York.

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

MUTCD (1988) *Manual on Uniform Traffic Control Devices*. U.S. DOT, Government Printing Office, Washington, D.C.

Newell G.F. (unpublished) Notes on transportation operations. Univ. of California, Berkeley, U.S.A.

Newell G.F. (1962) Theories of instability in dense highway traffic. *J. Opn. Res. Soc. Japan* ***5***, 9-54.

Newell G.F. (1971) *Applications of Queueing Theory*. Chapman Hall, London.

Newell G.F. (1979) Airport capacity and delays. *Transpn Sci.* ***13***, 201-241.

Newell G.F. (1989) Theory of highway traffic signals. Institute of Transportation Studies UCB-ITS-CN-89-1, Univ. of California, Berkeley, U.S.A.

Newell G.F. (1993) A simplified theory of kinematic waves in highway traffic *I:* General theory. *II*: Queueing at freeway bottlenecks. *III*: Multi-destination flows. *Transpn Res.* ***27B***, 281-313.

Newell G.F. (1995) Theory of highway traffic flow 1945 to 1965. Institute of Transportation Studies, Special Report, Univ. of California, Berkeley, U.S.A.

Newman L. (1986) Freeway Operations Analysis Course Notes. Institute of Transportation Studies, University Extension, Univ. of California, Berkeley, U.S.A.

Payne H.J. (1971) Models of freeway traffic control. *Proc. Math. Models Publ. Sys. Simul. Council* ***28***, 51-61.

Phillips W.F. (1977) Kinetic theory for traffic flow. Research Report for U.S. DOT. Logan, UT: Dept. of Mech. Eng., Utah State Univ., U.S.A.

Pitstick M. (1990) Measuring delay and simulating performance at isolated signalized intersections using cumulative curves. *Transpn Res. Rec.* ***1287***, 34-91.

Prigogine I (1961) A Boltzman-like approach to the statistical theory of traffic flow. *Proc. Int. Symp. on the Theory of Traffic Flow* (R. Herman, ed.), pp. 158-164, Elsevier, New York.

Prigogine I. and Herman R. (1971) *Kinetic theory of vehicular traffic*. Elsevier, New York.

Richards P.I. (1956) Shock waves on the highway. *Opns. Res.,* ***4***, 42-51.

Smilowitz K.R., Daganzo C.F., Cassidy M.J. and Bertini R.L. (1998) Some observations of traffic in long queues. Institute of Transportation Studies UCB-ITS-RR-98-6, Univ. of California, Berkeley, U.S.A. *to be published in Transpn Res. Rec.*

Solberg P. and Oppenlander J. (1966) Lag and gap acceptance at stop-controlled intersections. *Highway Res. Rec.* ***118***, 48-67.

TRB (1994) Special Report 209, *Highway Capacity Manual*. Transportation Research Board, Washington, D.C.

Vaughn R. and Hurdle, V.F. (1992) A theory of traffic flow for congested conditions on urban arterial streets. *I*: Theoretical development. *Transpn Res.* ***26B***, 381-396.

Vaughn R., Hurdle, V.F. and Hauer E. (1984) A traffic flow model with time-dependent o-d patterns. *Proc. Symp. on Transportation and Traffic Theory*, (J. Volmuller and R. Hamerslag, eds.), pp. 155-178, VNU Science Press, Utrecht, The Netherlands.

Webster F.V. (1958) Traffic signal settings. Road Research Technical Paper, No. 39, Road Research Laboratory, Ministry of Transport, HMSO, London.

Windover J.R. (1998) Empirical studies of the dynamic features of freeway traffic. *PhD thesis*. Univ. of California, Berkeley, U.S.A.